SELF-REPRESENTATIONALISM AND THE RUSSELLIAN IGNORANCE HYPOTHESIS:

A HYBRID RESPONSE TO THE PROBLEM OF CONSCIOUSNESS

THOMAS WILLIAM MCCLELLAND

DPHIL IN PHILOSOPHY

UNIVERSITY OF SUSSEX

2012

Table of Contents

INTRODUCTION

This thesis aims to provide a compelling and distinctive response to the Problem of Consciousness. This is achieved by offering a bipartite analysis of the epistemic gap at the heart of that problem, and by building upon the hypothesis that the apparent problem is symptomatic of our limited conception of the physical.

Chapter 1 introduces the problem. The key question is whether phenomenal consciousness is onticly dependent on the physical, or onticly independent of it. There are powerful arguments for the Primitivist view that consciousness is independent of the physical. These arguments rest on the apparent epistemic gap between the physical and the phenomenal. I propose that this apparent gap must be understood as a composite of two deeper conceptual gaps pertaining to the subjective character and qualitative character of consciousness respectively. The '*–tivity gap'* claims that physical states are objective, phenomenal states are subjective and that there is no entailment from the objective to the subjective. The '*–trinsicality gap*' claims that physical properties are extrinsic (structural), that phenomenal qualities are intrinsic (non-structural) and that there is no entailment from the extrinsic to the intrinsic. After refining the case for Primitivism, I consider the compelling reasons for *rejecting* Primitivism in favour of Physicalism. The challenge posed by the Problem of Consciousness is to resolve this antinomy between Primitivism and Physicalism.

In Chapter 2 I consider standard responses to the problem. The failings of these positions lead me to introduce three criteria that an adequate response must satisfy. I reject the view that Primitivism can be salvaged, and hold that a satisfactory response to the problem must protect Physicalism. I reject standard 'Type-A' responses according to which there is no epistemic gap between the physical and the phenomenal, and argue that a satisfactory response cannot deny the manifest reality of phenomenal consciousness. Finally, I reject 'Type-B' responses according to which the epistemic gap does not entail ontic distinctness. I hold that if Physicalism is true, the entailment from the physical facts to the phenomenal facts must be knowable a priori for an epistemically ideal subject.

Chapter 3 evaluates a non-standard Type-A response to the Problem of Consciousness which promises to satisfy all three criteria. According to Stoljar's Epistemic View (EV), consciousness only *seems* inexplicable in physical terms because we have a limited conception of the physical. I argue that EV should be supported iff two demanding challenges can be met: the Relevance Condition requires adequate reason to believe that unknown physical properties could address the –tivity gap and the –trinsicality gap. The Integration Condition requires adequate reason to believe that there is a specific blind-spot in our current conception of the physical that is plausibly occupied by properties that perform the requisite explanatory role. To satisfy these conditions, the advocate of EV must make positive claims about the content of our proposed ignorance.

In Chapter 4 I argue that EV stands or falls with the plausibility of the Russellian Ignorance Hypothesis (RIH). According to RIH, we have no concepts of the intrinsic properties of physical entities, and those intrinsic properties are integral to the physical explanation of consciousness. I argue that we are indeed conceptually ignorant of intrinsic physical properties. I also argue that RIH meets the Integration Condition, and goes some way to satisfying the Relevance Condition. RIH plausibly undermines the –trinsicality gap by showing that some physical properties are intrinsic, though they are beyond our current conception. The apparent gap is then an illusion resulting from the fact that all *known* physical properties are extrinsic. RIH fails, however, to address the –tivity gap. I conclude that no version of EV can offer a full response to the Problem of Consciousness.

In Chapter 5 I explore an entirely different kind of response to the Problem of Consciousness. Representationalism claims that consciousness is explicable in terms of intentional properties, and that intentional properties are explicable in terms of physical properties. I argue that standard Representationalist proposals are unable to account for the qualitative character of conscious states, and diagnose this failure in terms of the –trinsicality gap. However, the prospects for a Representationalist account of subjective character are more promising. Specifically, Kriegel's Self-Representationalism holds that a mental state is a phenomenal state in virtue of

suitably representing itself. I argue that this proposal plausibly addresses the –tivity gap.

RIH and Self-Representationalism each deal with one of the two apparent conceptual gaps between the physical and the phenomenal, but not the other. In Chapter 6 I develop a *hybrid* proposal that combines the best of both positions. The 'Neo-Russellian Ignorance Hypothesis' (NRIH) claims that a mental state is a phenomenal state at all in virtue of suitably representing itself, and has its qualitative character in virtue of the intrinsic physical properties involved in its implementation. I expand this claim and defend it against a number of potential criticisms. I also explore the relationship between its two components, suggesting that they are each founded on a common epistemic insight. I argue that NRIH successfully addresses the –tivity and –trinsicality gaps and, moreover, that it provides a compelling account of why consciousness appears to be inexplicable in physical terms. I conclude that NRIH offers a powerful response to the Problem of Consciousness that successfully undermines the case for Primitivism. Furthermore, I conclude that NRIH has substantial advantages over competing attempted responses, and offers the best possible way of capitalising on the insights of EV and Representationalism.

CHAPTER 1

THE PROBLEM OF CONSCIOUSNESS

The purpose of this chapter is to summarise the Problem of Consciousness. The problem arises in connection with the following question: what is the ontic relationship between consciousness and the physical? I will outline the two possible answers to this question: Primitivism, which claims consciousness is onticly distinct from the physical world, and Physicalism, which denies that claim. I consider the standard case in favour of Primitivism, examining the Conceivability Argument and the Knowledge Argument. The limitations of these arguments, however, lead me to supplement them with two further considerations, which I label the '*–tivity gap'* and the '*–trinsicality gap'*. I argue that these two conceptual gaps have important ramifications for the debate between Primitivists and Physicalists. After concluding that a serious case can be made in favour of Primitivism, I move on to consider the case *against* Primitivism. Focusing on the threat of 'epiphenomenalism', I conclude that a strong case can be made against Primitivism. This puts us in a position to formulate the Problem of Consciousness: when faced with the question of the ontic status of consciousness, we find compelling reasons to adopt a Primitivist stance, *and* compelling reasons to reject such a stance. Solutions to the problem must offer us a plausible resolution of this antinomy. In Chapter 2 I move on to consider the standard attempts to provide such a solution, and show why they are unsatisfactory.

SECTION 1

THE QUESTION OF CONSCIOUSNESS

What is the ontic relationship between consciousness and the physical world? This is the central question of the philosophy of consciousness, and any attempt to answer it faces the Problem of Consciousness. The subject matter of this question has a three-

part structure: the two relata and their ontic relation, each of which I will explore in this section. Setting up this question raises a variety of philosophical issues before we even consider how best to answer it. Clarifying concepts of consciousness and the physical is a philosophical task in its own right. In fact, proposed answers to the question often involve a distinctive analysis of these categories. Furthermore, asking about their ontic relationship presupposes some range of possible relations, but the nature of such relations is subject to a great deal of debate. Again, proposed answers to the question are often based on distinctive accounts of the possible relations. I will aim to offer a path through these issues that allows us to form a clear conception of the question without digressing too far into these complications.

1.1. CONSCIOUSNESS

The term 'consciousness' is notoriously ambiguous, and our target question concerns one specific aspect or variety of consciousness. To uncover the relevant sense of the term, we should first draw a distinction between two alternative ways of characterising mental states, including conscious states. A mental state can be described *functionally* or described *phenomenally*.[1] To describe a mental state functionally is to describe what it *does*. For example, the mental state of being in pain has a distinctive causal profile. It is the kind of state typically caused by bodily damage, and which typically has the effect of promoting avoidance behaviour. It will also make a range of standard contributions to the wider behavioural dispositions of its bearer, such as the disposition to say "I am in pain" when asked. Functional descriptions of a mental state are available from a third-person perspective: the causes and effects of a mental state can be observed by others. There are some uses of 'conscious' that can be defined in functional terms. For example, on Block's (2002) notion of 'access consciousness' a mental state is conscious when it is available to a subject for use in reasoning and action control. Other related uses of 'conscious' revolve around

[1] Güzeldere (1997, p.11) offers a particularly useful exposition of this distinction and its ramifications.

wakefulness, attention or verbal reportability. On all these accounts, if a state *does* what a conscious state does, then it *is* a conscious state.

To describe a state phenomenally is to describe how things *seem* for the subject of that mental state. Nagel (1974), building on Sprigge (1971), famously describes conscious experience in terms of 'what it is like' for the subject to be in that mental state. A state is *phenomenally* conscious iff there is something it's like to be in that state for its subject. The functional description of the pain state does not capture that the pain is being experienced, nor does it capture how that pain experience feels for its subject. To describe the phenomenal aspect of a mental state is to describe it from a first person perspective: to characterise how it seems for its bearer rather than how it manifests to an outside observer. Phenomenal states presumably have some functional profile, and it is plausible that a state's phenomenal and functional properties are intimately connected. Nevertheless, there is a clear conceptual distinction to be made between a state being phenomenally conscious and it having some particular functional profile. Mental states that are not phenomenal are plausibly open to a purely functional characterisation, but the concept of phenomenal consciousness is not a functional concept. The target question concerns phenomenal consciousness as opposed to any purely functional notion of consciousness. The functional notions raise a variety of interesting puzzles but, as Chalmers (1995/1997) argues, these are 'easy' problems in that we understand how it is possible for insentient matter to perform some functional role. By contrast, when we start to think about how the brain generates phenomenal consciousness, we run into the 'hard problem'.

What more can we say about the nature of phenomenal consciousness? Block proposes that '…all one can do is *point* to the phenomenon…' since there is no non-circular way of defining it (2002, p.206). In discussing phenomenal consciousness, we can only assume that we are each 'pointing' to the same type of state. This is an important claim, and we will often have recourse to appeal to the immediate and inarticulate grasp we have of what experiences are. That said, there is room for us to shed light on phenomenal consciousness by differentiating two aspects of phenomenal states; their *subjectivity* and their *qualitative character*.

Kriegel (2008, pp.45-57), drawing on Levine (2001), distinguishes between the subjective character and the qualitative character of phenomenal states. Subjectivity is the *existence condition* of a phenomenal state – the property in virtue of which it is a phenomenal state at all. A state is subjective iff there's something it's like to be in that state for its subject.[2] The pain caused by stubbing your toe is not just an event that occurs *in* you – it is painful *for* you, the subject. It is *your* pain. Similarly, a visual experience is reddish *for you*. An unconscious process associated with toe damage or sensitivity to redness is not presented to any subject – it is not experienced by anyone – so it is not a phenomenal state. There is a conscious experience iff there is a conscious subject experiencing it.[3] This characteristic of awareness, or seeming, or presentedness, is what distinguishes the conscious from the non-conscious.

Given that subjectivity is the existence condition of phenomenal states, all phenomenal states are subjective states. In what respect, then, can phenomenal states differ from one another? The *identity condition* of a phenomenal state – the feature that makes it the kind of phenomenal state it is – is its *qualitative character*. A state is phenomenal iff there's *something* it's like to be in that state, and its qualitative character constitutes *what* it's like to be in that state. An experience of pain and an experience of redness share the property of *subjectivity*, but they differ *qualitatively*. What it's like to undergo a pain experience is very different to what it's like to undergo a reddish experience. Phenomenal qualities are those fully specific properties that characterise our conscious lives. All experiential states have qualitative character. Just as something cannot have the determinable property 'being a shape' without having some determinate shape such as 'being a square', so a state cannot have the property 'being phenomenal' without having some determinate phenomenal character such as 'being reddish'. As this appeal to the 'determinable-determinate' relation indicates, the distinction between subjectivity and phenomenal character does not entail that

[2] We must be sensitive to the fact that there are other senses of the term 'subjective' in play in the consciousness literature, and that these different uses are often not distinguished with sufficient clarity. The use defined here is not intended to match all other uses.

[3] Furthermore, it is plausible that there must be *no more than one* subject of an experience. Your pain, or your impression of redness, is essentially yours and yours alone. The *privacy* of experience is bound to its subjectivity. However, it would detract from the pivotal claims about subjectivity to take on any commitments regarding privacy at this stage. See Strawson (2008) for an exposition of this.

phenomenal qualities and the awareness of those qualities are separate states.[4] Rather, qualitative character is the determination of a single state of awareness.

Of course, our experiential state at any given time will be characterised by a vast array of qualities. There are qualities distinctive to our different sense modalities, to our emotional states and perhaps to our intellectual states, all simultaneously contributing to our experience.[5] For instance, what it's like for a subject as they look at a painting might involve a visual impression of its colours, an emotional sense of admiration for the painter and an intellectual experience of thinking about its composition. The qualitative character of our experience at any given time is the sum of the qualitative properties it instantiates (Kriegel, 2008, p.46). Two experiential states are qualitatively identical iff what it's like for their subjects to be in those states is precisely the same.[6] As such, wherever two states differ in what it is like to undergo them, there must be some difference in qualitative character (see Stoljar 2005, p.469).

We thus have a characterisation of the first element of the question of consciousness. There will be more to say about the concept of phenomenal consciousness in due course. Nevertheless, we have enough of a notion of phenomenal consciousness to capture the target question. From this point forward, 'consciousness' will mean *phenomenal* consciousness. In this sense, a state is conscious iff it is a state of subjective qualitative awareness. Sometimes I will talk of a creature or person being conscious, rather than a state being conscious, but a creature or person is conscious at a time precisely if it is a bearer of conscious *states* at that time (Kriegel, 2008, p.28).

1.2. THE PHYSICAL

When asking about the relationship between consciousness and the physical, how

[4] In Chapter 5, I will argue that there are good reasons to reject such a claim of separability.

[5] The notion of qualities of intellectual experience is controversial. Horgan & Tienson (2002), Strawson (2008, pp.291-295) and others argue for the existence of such properties. Since the Problem of Consciousness can be captured using paradigm phenomenal properties such as pain and redness, we can sidestep this debate.

[6] This is compatible with the possibility that, in reality, no two experiential states have ever been precisely alike.

should 'the physical' be understood? One common route is to characterise physical entities in terms of physical theory. Rather than giving an a priori definition of what physical entities are, this approach defers to science to tell us about the physical: the physical is whatever physical theory tells us it is.[7] 'Physical theory' is typically understood in terms of fundamental physics, which describes the basic physical entities out of which all physical things are constituted.

To evaluate an account of 'the physical', we must consider its implications for the notion of 'Physicalism'. Physicalism is the view that all concrete entities are exhaustively constituted by fundamental physical entities. To ask what the ontic relationship is between the physical and the phenomenal is effectively to ask whether or not the existence of consciousness is compatible with a Physicalist ontology. I argue that the 'physical theory' account of what it is to be physical has unacceptable implications for the content of Physicalism.

The physical theory approach faces a problem known as 'Hempel's Dilemma'.[8] Should we define 'physical' in terms of current physics or in terms of a hypothetical complete physics? The first choice gives 'physical' a clear meaning, but renders Physicalism obviously false. It is completely implausible that the entities described by current physics constitute all concrete entities, including phenomenal states. The second choice sounds like it could be true. The problem, though, is that the meaning of 'physical' (and so of 'Physicalism') becomes obscure since we do not know what the complete physical theory looks like. We would have no substantive grasp on what Physicalism about consciousness really is, and so no good reason to ask whether or not it is true. Furthermore, this characterisation risks rendering Physicalism trivially true, since a physical theory is only complete if it succeeds in accommodating *all* concrete entities, which inevitably includes all phenomenal states.

How, then, should we characterise 'the physical' if not in terms of physical theory? Rather than offering necessary and sufficient conditions for being physical, I suggest we make do with a *minimal* condition of physicality. Fundamental physical properties are *non-phenomenal* properties. We have already characterised what it is to

[7] Chomsky (2009) argues that the concept 'physical' places no a priori constraints on what kind of property could be countenanced as physical by our best physical theories.
[8] See Crane & Mellor (1990), Levine (2001) and Stoljar (2010).

be phenomenal, so can give a straightforward negative characterisation of the non-phenomenal. Being non-phenomenal is not *sufficient* for being physical, but it is an important *necessary* condition.[9] On this account, Physicalism about the phenomenal is true only if phenomenal properties are ultimately realised by properties that are not themselves phenomenal. Physical theory still has a role to play: without defining the physical in terms of physical theory, science remains our guide to how the physical world is.

Maybe a richer account of physicality is available – an account that goes beyond the negative characterisation and perhaps overcomes Hempel's dilemma to bind physicality to physical theory. Even if this were so, it remains the case that the minimal condition is the most appropriate way of capturing the question of consciousness. There is an intuitive puzzle concerning whether or not consciousness boils down to anything more simple then itself. This puzzle does not arise due to any complex characterisation of *what* consciousness might boil down to, so to define 'the physical' as anything more than non-phenomenal would distort the driving question, and take us into controversial territory unnecessarily.[10]

1.3. ONTIC RELATIONS

We now have our two *relata*, the phenomenal and the physical, but we are yet to explore the kind of *relation* with which the question of consciousness is concerned. Any two categories of state will stand in any number of relations to one another, but we are concerned specifically with the *ontic* relationship between the physical and the phenomenal. What kind of dependence, if any, does the existence of one category of state have on the other? Perhaps we could form a complete list of the candidate relations, and so see the range of possible answers to the question. The difficulty here is that there is no consensus on what the options are. A wide range of relations have

[9] I will extend the conditions of physicality further in Section 4.1.

[10] Others who advocate something like this minimal account include Montero (1999), Levine (2001), and Spurrett & Papineau (1999), though they have come under criticism by Judisch (2008) and others. Stoljar (2006) and McGinn (2004, especially pp.18-19) make similar points but choose to remove the term 'physical' from discussion, though this is more a terminological difference than a difference in position.

been proposed, but some have been dismissed as meaningless, some have been claimed to collapse into one another and others have been accused of being too broad to constitute an informative answer to the question at hand.[11]

At this stage we can side-step some of these issues by making the question more precise. Is the phenomenal onticly distinct from the physical, or onticly dependent upon it? That is, does the instantiation of consciousness involve something over and above the instantiation of any non-phenomenal properties, or is the instantiation of certain non-phenomenal properties sufficient for the instantiation of consciousness? Kripke (1980) offers a useful way of understanding this kind of question: if God created the universe, once he set how things are physically, did he thereby fix how things are phenomenally, or did he have to create the phenomenal separately? If how things are phenomenally was already fixed, then the phenomenal is *not* onticly distinct from the physical. If a further creative act was required, then the phenomenal *is* onticly distinct from the physical.

Ontic dependence should not be confused with causal dependence. Consider an image on a computer screen, the pixels that compose that image, and the computer to which that screen is hooked up. The computer screen's state of displaying a particular image is *causally* dependent on some state of the computer. But this is not a case of ontic dependence. The two states are separable – it is possible for the image on the screen to remain in the absence of that state of the computer, or in the total absence of the computer. By contrast, the computer screen's state of displaying a particular image is *onticly* dependent on the state of the screen's pixels. The two states are inseparable. The existence of the image depends on the existence of the pixels. To describe this as an identity relation would be too strong since the two states could have different properties. It is better to describe it in the following terms: the state of the screen is *nothing more than* the state of the pixels. The existence of the image is exhausted by the existence of the pixel state. There is no further ingredient involved in the occurrence of the image.

Are phenomenal states nothing more than physical states, or are they

[11] Options presented include emergence (Broad 1925), weak and strong supervenience (Kim 1982) superdupervenience (Horgan 1993), reduction (Churchland 1996) and identity (Smart 1959).

separable from one another? As this formulation shows, the question is not one of identity. To claim that any phenomenal state is identical to a non-phenomenal state would be incoherent since the two states will differ from each other in at least one respect: one is phenomenal and the other is non-phenomenal. By contrast, to claim that a phenomenal state is *exhaustively constituted* by non-phenomenal goings-on is far from being plainly incoherent. It might be no more problematic than an image being constituted by things that are not images. As such, the interesting question of ontic dependence should not be confused with the un-interesting question of identity.

The question of ontic dependence can informatively be put in modal terms. Is there a possible world in which the A-property instantiations remain exactly as they are, but the B-property instantiations differ? More precisely, is there a 'minimal A-property duplicate' of this world in which the A-properties are held constant and no extra properties are added, but in which the B-properties differ?[12] If so, the B-properties are not onticly dependent on the A-properties. The B-properties are *further properties* that contribute to how the world is. Once the A-properties have been set, it remains an open question what the distribution of B-properties is, so the B-properties are something *over and above* the A-properties. If, by contrast, any world where the A-properties are such is also a world where the B-properties are such, then the B-properties are onticly dependent on the A-properties. That is, the A-properties *necessitate* the B-properties. This account of dependence will apply *mutatis mutandis* to talk of states, objects and entities.

Applying the schema to consciousness, is there a possible world that is a minimal physical duplicate of the actual world, but in which how things are phenomenally differs from the actual world? If so, the phenomenal is onticly independent of the physical. If not, the phenomenal is onticly dependent on the physical. There are many different *ways* of being onticly dependent, but identifying these different ways would lead us into unnecessarily complicated territory. Ontic (in)dependence is a relatively intuitive notion; after all, it is only by appealing to that

[12] This 'minimal duplicate' clause is required so we can ignore worlds in which the A-facts are the same as in the actual world but some extra factor means the B-facts differ. This world would differ from ours with respect to the B-facts, but should not count against the ontic dependence of B-facts on A-facts in *our* world.

intuitive notion that philosophers have tested the accuracy of proposed accounts of ontic (in)dependence. It is this notion that is key to the question of consciousness. We are now in a position to formulate the target question as follows:

> ***The Question of Consciousness****:* Is the phenomenal onticly dependent on the physical, or onticly independent of the physical?

Primitivism is the view that the phenomenal is onticly independent; that phenomenal properties are basic non-physical features of reality. Physicalism is the view that the phenomenal is onticly dependent on the physical; that phenomenal states are not primitive components of the world, but are rather necessitated by how things stand physically. Our minimal understanding of the term 'physical' brings with it a minimal understanding of the term 'Physicalism'. Others may insist on reserving the label for a stronger position – perhaps involving reducibility to the terms of fundamental physics – but this is not how I will use the term. For us, the core commitment of Physicalism is that phenomenal properties are onticly dependent on non-phenomenal properties.

SECTION 2

THE INITIAL CASE FOR PRIMITIVISM

There are two main arguments in favour of Primitivism: the Conceivability Argument (CA) and the Knowledge Argument (KA). CA and KA adopt the same general strategy. I begin by considering the shared structure behind those two arguments.[13] Identifying that structure will allow us to better appreciate how those arguments work, what the relationship is between them, and where a critic can raise objections relevant to both arguments. I then move on to outline each argument separately. In Section 3, though, I will suggest that in order to make the best possible case for Primitivism these initial arguments must be supplemented by further arguments.

[13] This simple shared structure is explained by Chalmers (2002).

2.1. THE PRIMITIVIST STRATEGY

The Schematic Argument (SA) for Primitivism runs as follows:

> SA1) There is an epistemic gap between the physical and the phenomenal.
>
> SA2) If there is an epistemic gap between the physical and the phenomenal, then there is an ontic gap.
>
> SA3) Therefore, there is an ontic gap between the physical and the phenomenal.

2.1.1. The Epistemic Step

SA1 is the epistemic step of an argument for Primitivism. This step must establish that there is failure of epistemic entailment between the physical facts and the phenomenal facts. No knowledge of the physical facts could ever *explain* the phenomenal facts (on an appropriately strong understanding of 'explain'). Levine (2002) labels this the 'explanatory gap'. The kind of entailment in question is a priori entailment. Call the totality of physical facts 'P' and the totality of phenomenal facts 'Q'. There is an epistemic entailment from P to Q iff the conditional proposition 'P→Q' is knowable a priori. Of course, neither P nor Q can be known a priori: they are contingent complex facts. It is only the conditional proposition, often labelled 'the psychophysical conditional', that must be knowable a priori. The proposition 'George Clooney is a bachelor' is not knowable a priori, but the proposition 'if George Clooney is a bachelor then he is male' *is* knowable a priori. Knowledge that the antecedent holds would be a posteriori, but knowledge of the conditional as a whole is a priori. Contrast this with the conditional claim 'if George Clooney directs a film next year, he'll win an Oscar'. This proposition is *not* knowable a priori. Establishing the truth of the antecedent is again an a posteriori matter, but this time the conditional as a whole can only be established by looking to the world rather than by looking to the concepts it involves.

Primitivists deny that 'P→Q' is an a priori truth. If true, it is only an a posteriori truth. They claim that there is an 'epistemic gap' between the physical facts and the phenomenal facts. Various thought-experiments can be used that point to this

epistemic gap, but in and of themselves these will not generate any ontic conclusions. Establishing an epistemic gap would tell us something about the relationship between our physical and our phenomenal *concepts*, but further work is required if a conclusion is to be drawn about the relationship between physical and phenomenal *properties*. This is the job of the second step.

2.1.2. The Ontic Step

Primitivism holds that there is an *ontic* gap between the physical and the phenomenal. In order to reach this ontic conclusion from an epistemic premise, a conditional binding the two is required. If there is an epistemic gap between the physical and the phenomenal, then there is an ontic gap. This conditional claim is motivated by the more general thought that if there is an epistemic gap between A-facts and B-facts, then there is an ontological gap between them. It would be implausible to deny that this ontological conditional holds generally but to maintain that it holds in the special case of the physical and the phenomenal. The case for Primitivism rests on the wider claim that *any* failure of a priori entailment means a failure of ontic entailment.

If the phenomenal is onticly dependent on the physical, then 'P→Q' is a necessary truth. The ontic step claims that if 'P→Q' is a necessary truth, then 'P→Q' must be knowable a priori. By showing, in the epistemic step, that 'P→Q' is not knowable a priori, the Primitivist can thus infer that the phenomenal is onticly independent of the physical. Though this inference is clearly valid, the Primitivist must give us reason to accept the two steps. CA and KA take different routes to establishing that there is an epistemic gap. There are also various ways to establish the ontic step, but I will postpone discussion of these until Chapter 2.

2.2. THE CONCEIVABILITY ARGUMENT (CA)

2.2.1. Conceivability and Entailment

The strategy of CA is to use what we can conceive of as a test of epistemic entailment.

Conceivability and epistemic possibility come hand in hand, as do inconceivability and epistemic impossibility. Imagination is a testing ground for what our concepts can do. Roughly, if our concepts can formulate a certain scenario in our imagination, that scenario is an epistemic possibility. If our concepts cannot do so, then the scenario is epistemically impossible. Conceiving of a flying pig shows that such a creature is an epistemic possibility – there is nothing about the concepts 'pig' and 'flying' that makes a flying pig unthinkable. By contrast, round squares are not a genuine epistemic possibility and, accordingly, are inconceivable.

Of course, conceivability tests can misfire. Sometimes we think we are conceiving of one scenario when really we are conceiving of another. Say *p* is conceivable and *q* is inconceivable. If you conceived *p* but believed that you were conceiving *q*, you would mistakenly claim that *q* is an epistemic possibility. This is a case of 'proposition confusion' (Stoljar 2006, p.74). But this possibility does no damage to the claim that if we *really are* conceiving of a scenario, then it is epistemically possible.

A further consideration is that conceivability only entails epistemic possibility if it is the *right kind* of conceivability. Van Cleve distinguishes between strong and weak conceivability (see Stoljar, 2006, p.75). A subject weakly conceives of *p* if they entertain *p*, and it is not the case that *p* strikes them as impossible. A subject strongly conceives of *p* if it imaginatively appears to them that *p* is possible. Significant considerations show that weak conceivability is a poor test of epistemic possibility. A conceivability test might then misfire if we take ourselves to be *strongly* conceiving of *p*, when really we are only weakly conceiving of *p*. This mistake is known as 'mode confusion' (Stoljar 2006, p.75). By 'conceivable', we will always mean *strongly* conceivable unless otherwise stated.

What does epistemic possibility have to do with epistemic entailment? If '$A \rightarrow B$' is knowable a priori, then '$A \wedge \neg B$' must be a priori false. If '$A \wedge \neg B$' is a priori false, then '$A \wedge \neg B$' is epistemically impossible. It must be the kind of proposition that we can know is false just by reflecting on the concepts involved. From this it follows that if we can conceive of '$A \wedge \neg B$', then '$A \rightarrow B$' is not an a priori truth. Applying this to consciousness, the question becomes whether '$P \wedge \neg Q$' is conceivable. That is, can we

conceive of a scenario in which the physical facts are held exactly as they are in reality, but in which how things are phenomenally is different?

2.2.2. Zombies and Inverts

To perform an appropriate conceivability test, we should start with something more manageable than P, the complete set of physical facts, and Q, the complete set of phenomenal facts. A sub-set of those facts should do the job. Since we only have direct access to our own phenomenal states, the scenario we attempt to conceive should involve our conscious experiences. Accordingly, the non-phenomenal facts in question should be those pertaining to our own physico-functional constitution.[14] Can we conceive of a physical duplicate of ourselves – a being like us in all non-phenomenal respects – but who differs from us phenomenally? To answer this, we should consider some alternative ways in which this duplicate might 'differ from us phenomenally'. For our purposes, two types of phenomenally divergent physical duplicates will be informative: zombies and inverts.

We have phenomenal consciousness. A being with *no* phenomenal states therefore differs from us in a phenomenal respect: there are phenomenal states that we have and they do not. One way of imagining a duplicate like you in all non-phenomenal respects, but unlike you phenomenally, is to imagine your *zombie twin*. Your zombie twin has *all* the same physical characteristics you have, but has no conscious states. The notion of 'zombie twins' is championed by Chalmers (1996). The Primitivist cannot *show* that zombie twins are conceivable. They can only ask you to perform the conceivability test. Many claim to find their zombie twin conceivable, so CA has some serious purchase here.

We have phenomenal states with a particular qualitative character. A being that has phenomenal states with a different qualitative character would differ from us phenomenally. If we are to conceive of a being like us in all physical respects, but who differs from us in respect to their qualitative character, it is useful to have an idea of what qualitative character their experience has. Shoemaker (1982) introduces the idea

[14] One might insist that the physical states responsible for consciousness extend beyond the individual. If so, we can simply adjust the conceivability test to incorporate those wider physical states.

of 'qualia inversion'. Our visual experiences are characterised by a rich spectrum of colour-qualities. A qualia invert is someone whose colour spectrum is turned upside down relative to ours. The quality we enjoy when looking at green objects, they enjoy when looking at red objects, and vice versa. They are responsive to all the same visual properties as we are, but have different experiences when presented with those properties. Your *invert twin* has all the same physical characteristics you have, but their colour-qualities are inverted relative to your own. Again, such a being is widely held to be conceivable.

There are many other ways of imagining beings like us physically, but unlike us phenomenally.[15] Why focus on precisely these two? Consider the distinction drawn earlier between subjectivity – the existence condition of a phenomenal state – and qualitative character – the identity condition of a phenomenal state. A zombie twin is a being devoid of subjectivity, where an invert twin is a being who has subjective awareness, but is such that the qualitative character of that awareness diverges from our own. Later we will see that worries about the explanation of subjectivity, and worries about the explanation of qualitative character, can come apart. As such, it will be useful to have conceivability scenarios that address each aspect of consciousness separately. All other available conceivability scenarios are simply variations on these two: different ways of changing whether the duplicate has subjective awareness, or different ways of changing the character of that awareness. As such, they will not add anything substantial to our inquiry.

2.2.3. Conceivability to Possibility

Zombies and inverts are conceivable, and therefore epistemically possible. But are they *metaphysically* possible? Perhaps our physical *concepts* fail to entail anything about the phenomenal, but physical *properties* actually necessitate the instantiation of phenomenal properties. This is where the ontic step comes in. The claim is that there is no such thing as ontic dependence without some kind of conceptual entailment. Consequently, the conceivability of zombies and inverts shows that they are

[15] For a review of the various proposed conceivability scenarios see Stoljar (2006, pp.37-38).

metaphysically possible. If zombies and inverts are possible, the phenomenal is not necessitated by the physical, therefore Primitivism is true. We will leave discussion of whether this move from conceivability to possibility is defensible until later when we consider critics who reject that move. The general Conceivability Argument (CA) then goes as follows:

> CA1) A being identical to you in all physical respects, but which differs from you phenomenally, is conceivable.
>
> CA2) If such a being is conceivable, then phenomenal states are not epistemically entailed by physical states.
>
> CA3) If phenomenal states are not epistemically entailed by physical states, then they are onticly independent of the physical.
>
> CA4) Therefore the phenomenal is onticly independent of the physical.

The 'zombie argument' is the same as CA, but with the phrase 'differs from you phenomenally' in CA1 replaced with the more specific 'has no phenomenal consciousness'. Similarly, we can take the 'invert argument' to be the same as CA, but with that phrase replaced with 'has spectrum-inverted qualia relative to your own'.

2.3. THE KNOWLEDGE ARGUMENT (KA)

2.3.1. Mary the Neurologist

An interesting way of exploring whether there is an epistemic entailment between the physical and the phenomenal is to consider a subject who has *complete* relevant knowledge of the physical. Jackson (1982) invites us to imagine Mary the neurologist.[16] Mary has been confined since birth to a black and white room. In this room, she learns everything there is to know about the science of colour, including the physics of colours and the neurophysiology of colour perception. She has *complete* knowledge of the physical facts associated with seeing colours, but she has never seen colour herself. One day, she escapes her monochromatic prison and stumbles across a ripe tomato. For the first time, she experiences redness. Clearly, Mary learns something new here that her science textbooks could not tell her. She learns *what it's like* to

[16] Farrell (1950) offers an important pre-cursor to Jackson's argument.

experience redness.

What bearing does Mary's discovery have on whether there is an epistemic entailment from the physical to the phenomenal? What it's like to experience redness is a phenomenal fact. If that fact was epistemically entailed by the physical facts, Mary would have known it *before* she escaped her room. She had knowledge of those physical facts and, we can stipulate, has an unlimited ability to extrapolate the conceptual implications of that knowledge. Since Mary learns something new on seeing the tomato, what she learns cannot have been epistemically entailed by the physical facts.

There are many controversies surrounding what this scenario really shows, but one in particular is worth mentioning. The above presupposes that if *p* epistemically entails *q*, then someone with full knowledge of *p* can infer that *q*. This might usually be the case, but there is something unusual about the Mary scenario that puts pressure on that generalisation. Mary had no *concept* of phenomenal redness before leaving her room. It is too strong to say that *p* epistemically entails *q* only if knowing *p* automatically provides one with the concepts required to entertain the proposition *q*.

To get round this complication, we need to imagine a less neat, but more revealing scenario. Stoljar (2005) introduces *experienced* Mary, who has the same life-story described above, only at some point before her escape she is kidnapped. The kidnappers show her something red, then give her a pill that makes her forget what she saw and when she saw it, before returning her to the monochromatic room. Now Mary has a *concept* of phenomenal redness, but no knowledge connecting it with any of her physical knowledge. On escaping the room, experienced Mary still learns something new. She learns that her qualitative concept applies to what it is like to perceive red (and not, say, to perceptions of green). If the phenomenal facts were epistemically entailed by the physical facts, Mary would already have known this and so would have nothing new to learn. From here on, by 'Mary' I will mean *experienced* Mary.

So far we have stayed firmly on the epistemic level. Of course, for the Mary scenario to lend support to Primitivism, the ontic step must be added. If Mary cannot infer the phenomenal facts from the physical facts, then there is no ontic entailment

from the physical to the phenomenal. What Mary learns is a new phenomenal fact. We can formulate KA as follows:

> KA1) Mary knows all the physical facts before leaving her room, and learns a phenomenal fact on leaving her room.
>
> KA2) If Mary learns a phenomenal fact on leaving her room, then the phenomenal facts are not epistemically entailed by the physical facts.
>
> KA3) If the phenomenal facts are not epistemically entailed by the physical facts, then they are not onticly entailed by the physical facts.
>
> KA4) Therefore the phenomenal is not onticly dependent on the physical.[17]

2.3.2. KA's Relationship with CA

A brief comparison with CA is in order. The real difference between CA and KA is in their first premises: each argument offers a different way of establishing the epistemic gap, and then does the same thing with that gap to establish Primitivism. Some find the concept of *conceivability* problematic, and KA has the advantage of deploying the more straightforward notion of *learning something new*.

With CA, the zombie and invert formulations homed in on the subjectivity of consciousness and the qualitative character of consciousness respectively. As it stands, KA is only concerned with qualitative character – with Mary's discovery of *what it is like* to experience red. It is hard to see how a parallel scenario could be formulated in which Mary is ignorant of subjective awareness as such. Purging a particular kind of subjective state from Mary's life – the reddish kind – is one thing, but purging all subjective awareness is quite another. We would have to imagine a *zombie Mary* who studies the complete science of subjective awareness without ever having a phenomenal state.[18] One day she has her first conscious experience, and learns that *this* is the special kind of state associated with all those non-phenomenal states she had been studying. It is hard to make sense of Mary being a subject of knowledge at all before she becomes conscious, or at least hard to understand it as the same subject

[17] I take it as given that the transition from 'not being onticly entailed by the physical facts' to 'not being onticly dependent on the physical' is innocuous.

[18] McGeer (2003) introduces a 'zombie Mary', but the scenario she describes and the use it is put to are not what we are after.

who is both the zombie at one time then the conscious subject at another. Furthermore, there is unlikely to be a viable analogue of *experienced* Mary in this scenario: a Mary who has a full concept of what subjective awareness is, but who has forgotten which physical states it is associated with. We should conclude that KA is a vivid way of capturing the failure of epistemic entailment between physical facts and the facts of phenomenal character, but should rely on the zombie argument to capture the gap between the physical and the subjective-as-such.

SECTION 3

THE REFINED CASE FOR PRIMITIVISM

We have outlined the standard case for Primitivism in terms of CA and KA. The purpose of this section is to improve upon this standard case. I will present a *rudimentary response* to CA and KA. Though it is clear that this response is inadequate as it stands, it is important to understand *why* it fails. I argue that to fend off the rudimentary response, Primitivism must appeal to two conceptual gaps that I dub the '–tivity gap' and the '–trinsicality gap'. Here I draw on what Primitivists have actually said on the subject, but also make the more idiosyncratic claim that these two gaps require us to re-think the status of CA and KA. I conclude that the refined case offers serious reasons to advocate Primitivism.

3.1. THE RUDIMENTARY RESPONSE TO PRIMITIVISM

When presented with the initial case for Primitivism, one straightforward response would be to claim that science will eventually close the epistemic gap between the physical and the phenomenal. After all, a failure of explanation does not constitute evidence of in-principle inexplicability. The claim is that though we do not presently have an explanation of phenomenal consciousness in physical terms, such an

explanation will eventually be uncovered. This thought can be reinforced with reference to the state of current science. Neuroscience and cognitive science are very young, so we have a limited grasp of the kind of explanations that they will ultimately be able to provide. Perhaps they just need to be given the chance to chip away at the task of explaining consciousness. One could also speculate that the epistemic gap will be closed by progress on some other frontier of science; for example, there is a great deal of speculation about quantum phenomena being integral to the explanation of consciousness (Penrose 1989, Lockwood 1989, Atmanspacher 2011).

How does this kind of response address CA and KA? Regarding CA, the claim is that when we try to imagine zombies or inverts, we are not really conceiving of complete physical duplicates of ourselves. We do not have the complete science *required* to conceive of a physical duplicate of ourselves in any real detail. Once our scientific understanding is fully developed, however, we will be able to conceive of complete physical duplicates of ourselves, but it will be *inconceivable* to us for them to differ from us phenomenally. Regarding KA, the claim is that we have no substantive grip on Mary's epistemic situation, as her complete physical knowledge will include scientific theories that we do not have at our disposal. Consequently, we are not in a position to conclude that she would learn anything new on escaping her monochromatic prison.

At this point, the Primitivist might insist that not only do we currently lack a scientific explanation of consciousness, we cannot even *imagine* what such an explanation would be like. This deeper sense of mystery is what distinguishes the 'hard problem' of consciousness from more mundane explanatory problems - problems that appear open to scientific solutions. This defence, however, is ineffective. P.S. Churchland emphasises that '[a]dding "I cannot imagine explaining *P*" merely adds a psychological fact about the speaker, from which again, nothing significant follows about the nature of the phenomenon in question.' (1996, p.407) She goes on to explain, '[g]iven that neuroscience is still very much in its early stages, it is actually not a very interesting fact that someone or other cannot imagine a certain kind of explanation of some brain phenomenon.' (1996, p.407) In other words, the fact that

we cannot imagine the proposed explanation of consciousness does not constitute evidence that there is no such explanation.

This attack on Primitivism can be supported with reference to other explanatory problems. 'Life' once appeared to be a property that could not be explained in more basic terms i.e. in *non-life* terms (Dennett 1996, p.4, P.S. Churchland 1996, p.407). Just as Primitivism posits basic phenomenal properties, Vitalism posits a basic life force. Thanks to the progress of science, though, life can now be explained in more basic terms. The Vitalists were simply wrong that there was a case of permanent inexplicability here, rather than a mere temporary failure of explanation. Just as the Vitalists should not have trusted the intuitions of inexplicability that they had in their limited epistemic position, the Primitivist should not trust the intuitions of inexplicability they have from their analogously limited epistemic position.

One way of understanding this challenge to Primitivism is through the distinction between the 'hard' and 'easy' problems of consciousness. We previously recognised that *psychological* consciousness raises a range of explanatory problems, but suggested that these were easy in comparison to the deeper problem of explaining *phenomenal* consciousness (Section 1.1.1). However, given that we do not yet have the scientific theories that resolve the so-called 'easy problems', how could we be in a position to know that discovering the full cognitive story of consciousness will not simultaneously solve the 'hard problem'?[19] How can we be sure that future theories will leave an unexplained residue, despite our being deeply ignorant of the content of those theories?

In Chapter 3, we will consider a more sophisticated version of this kind of attack on Primitivism. In the meantime, however, we should ask how Primitivism can defend itself against the relatively simple objection just outlined. Why should we believe that no discoveries about the physical world, such as might be provided by future brain science, could close the apparent epistemic gap between the physical and the phenomenal?

[19] See Dennett 1996, p.5.

The standard defence against this line of thought is to argue that *more of the same won't do* – that further developments in science will inevitably yield the *wrong kind* of information to explain the phenomenal (e.g. Chalmers, 2002 pp.258-9). Defenders of the epistemic gap concede, as they should, that we are a long way from being epistemically ideal subjects. Nevertheless, they maintain that even from our limited position, we have a grip on what *kind* of knowledge an ideal subject would have, and so have a grip on whether such knowledge could explain phenomenal states. Of course, the plausibility of this move depends upon how the notion of 'wrong kind' is fleshed out.

3.2. TWO CONCEPTUAL GAPS

If there is a principled 'conceptual gap' between the physical and the phenomenal, we can be sure that no future discoveries about the physical will allow the epistemic gap to be closed. There are two conceptual gaps that might do the job. In the long term I will attempt to undermine these gaps, but in the meantime I will try to capture their intuitive force. Interestingly, these two gaps map on to the two aspects of phenomenal states identified earlier: subjectivity and qualitative character. This will have significant implications for our understanding of the epistemic gap. Though the distinction between the two conceptual gaps and the distinction between the two aspects of consciousness are each fairly well recognised, this correspondence between them is not.[20]

3.2.1. The –tivity Gap

The –tivity gap, as I will call it, pertains to the *subjective aspect* of phenomenal states. Phenomenal states are subjective in that there is something it's like to be in a phenomenal state for its subject. There are other senses in which phenomenal states

[20] For instance, Feser (2001, p.3) suggests that conscious states involve two characteristics – subjectivity and intrinsicality – that are each problematic. He does not, however, describe their relationship with one another and with the physical world in the way I do.

might be called subjective, but they do not concern us here. Physical states are objective. They exist, but they do not exist *for a subject* in the way that subjective states do. An objective state need not be *presented* to anyone; it just *is*. The compelling thought that drives the –tivity gap is that there can be no entailment from the objective to the subjective. Facts about how things are cannot entail facts about things *seeming* some way to a subject. For any objective state, it is always an open possibility for that state not to be accompanied by any kind of awareness – for it to lack any inside view. Similarly, how a state is objectively will never rule out the possibility that there is something it's like to occupy that state. We can summarise this line of argument as follows:

> TIV1) All physical states are objective states.
>
> TIV2) All phenomenal states are subjective states.
>
> TIV3) There can be no epistemic entailment from the objective to the subjective.
>
> TIV4) Therefore, there can be no epistemic entailment from the physical to the phenomenal.

If this argument is taken seriously, it is not merely that the objective facts with which we are familiar are unsuited to the explanation of subjective awareness. Rather, objective facts are simply *the wrong kind of fact* to entail the existence of subjective states. Though a complete science may contain facts radically different from those we find in current science, it is plausible that these facts will still be exclusively objective (Nagel 1974, p.527). Levine claims that '[n]o matter how rich the information processing or the neuro-physiological story gets, it still seems quite coherent to imagine that all that should be going on without there being anything it's like to undergo the states in question.' (2002, p.359) Consequently, the speculation that future discoveries about the physical world will close the epistemic gap can be ruled out.

In our discussion of CA, we identified that the zombie scenario pertained to the subjectivity of consciousness while the invert scenario pertained to the qualitative character of consciousness. As such, the –tivity gap is best understood in connection to the zombie scenario. To conceive of a physical duplicate of yourself is to conceive of an *objective* duplicate of yourself i.e. a being like you in all respects that do not, in and of

themselves, involve any subjective awareness. If there is a principled conceptual gap between the objective and the subjective, we can be sure that no matter how much we fill in the details of our imagined objective duplicate, it will remain conceivable for that duplicate to be devoid of subjective awareness. Future discoveries will have no impact on the conceivability of zombies.

3.2.2. The –trinsicality Gap

The –trinsicality gap is based on the idea that phenomenal qualities are non-structural, or *intrinsic*, while physical properties are structural, or *extrinsic*. This gap pertains to the *qualitative aspect* of phenomenal states.[21] Physics characterises fundamental physical entities structurally. It describes the spatiotemporal structure of entities - how those entities are located in space-time. It also describes the causal properties of those entities. These causal properties '…are ultimately defined in terms of spaces of states that have a certain abstract structure…such that the states play a certain causal role with respect to other states.' (Chalmers 2002, p.258) In other words, the causal properties are powers to influence the location of entities in space-time, and their location in abstract state spaces. Physics thus describes a rich web of relations between entities but never describes any non-structural properties.[22]

Alter explains that '[n]ot only does current microphysics tend to characterise its basic properties in solely structural/dynamic terms: it is a reasonable, if controversial, conjecture that completed physics would so characterise all its basic properties' (2009, p.760). It looks like we can be sure that an ideal physics will describe the world in structural terms.[23] Furthermore, we have reason to believe that '…from structure and dynamics, one can infer only structure and dynamics.' (Chalmers 2002, p.259) Though physics might entail the facts of biology or of economics, these are plausibly still structural facts, just on a different scale to those described by microphysics. Structural properties can never entail non-structural properties.

[21] See, among others, Alter 2009 and Montero 2010 for related arguments.
[22] Though this point is best understood in terms of fundamental physics, it is equally relevant to brain sciences (see Bolender 2001, p.44).
[23] It might be noticed that this argument ignores the possibility of non-structural physical properties that aren't revealed by physical theory. I will take advantage of this loop-hole in due course.

The qualities that characterise phenomenal consciousness are non-structural properties. A reddish quality stands in many relations, and has a specific location in the quality-space of all colour qualities. To describe the quality in terms of its relational profile, however, would be to miss out its essential nature – its *redness*. If a subject experiences a reddish quality but is totally ignorant of the relational features of that quality, they would still be acquainted with *what redness is*. Conversely, building on the Mary scenario, a subject with complete structural knowledge who knows, for instance, everything about the location of reddish qualities in some perceptual state space, would not thereby know what reddish experiences are like. The –trinsicality gap supports the epistemic gap with the following argument:

> TRIN1) All physical properties are structural properties.
>
> TRIN2) All phenomenal states involve the instantiation of non-structural properties.
>
> TRIN3) There can be no epistemic entailment from the structural to the non-structural.
>
> TRIN4) Therefore, there can be no epistemic entailment from the physical to the phenomenal.

The claim is that structural facts are the wrong kind of fact to entail anything non-structural, and since any future insights into the physical world will be exclusively structural, the epistemic gap will inevitably remain untouched.[24] Inverts will remain conceivable, and it will still appear that Mary would learn something new. Overall, the principled conceptual gap between the structural and the non-structural indicates a principled conceptual gap between the physical and the phenomenal.

3.3. THE DIALECTICAL SITUATION

3.3.1. The Relationship of the Conceptual Gaps

The –tivity gap presents Physicalists with the following challenge: why is there something it's like to occupy a state with particular *objective* physical properties rather

[24] I will postpone discussion of how this fits in with the intrinsic-extrinsic dichotomy until Chapter 4, where we will also lend further support to 'TRIN1'.

than nothing at all? The –trinsicality gap presents them with a further challenge: why does what it's like to occupy a state with particular *structural* physical properties have the qualitative character it has rather than some other? These two questions overlap in a variety of ways, and are easily confused with one another, so it will be worth our while to clarify how exactly they are connected.

The –tivity gap claims that the subjectivity of conscious states cannot be accounted for in physical terms. It also claims that it is the *objectivity* of the physical that is responsible for this failure of entailment. Some defenders of the epistemic gap would affirm the former claim but deny the latter. Chalmers (1996/2002), for example, holds that it is the *structural* nature of the physical that makes an explanation of subjective awareness impossible. This would extend the –trinsicality gap so that it pertains to the subjectivity of consciousness rather than just to its qualitative character. I argue that such an extension is inappropriate.

Chalmers claims that '[f]or any complex macroscopic structural or dynamic description of a system, one can conceive of that description being instantiated without consciousness' (2002, p.259). Though this is plausibly correct, it is misleading. The implication is that it is because such a description is *structural* that it inevitably fails to entail the occurrence of subjective awareness. This suggests that if a *non*-structural description of physical states was available, the subjectivity of phenomenal consciousness would cease to present an explanatory obstacle. But why should we think this? So long as the non-structural description is objective – so long as it does not already involve an experiential point of view – it will remain conceivable for that description to be instantiated without consciousness. To show this, we should consider what a non-structural description of the physical world might involve.[25]

Imagine a world in which qualitative redness is a fundamental physical property. As we know from the –trinsicality gap, redness is a non-structural property. Levine argues:

[25] In Chalmers's defence, there is a sense in which subjective awareness itself is non-structural, and therefore inexplicable in structural terms. The plausibility of this claim depends on the sense of 'structural' in play. When we explore the notion of structure in greater depth in Chapter 4 (Section 4.4.), it will emerge that subjectivity is not relevantly non-structural.

> …if nature just has a richer stock of basic properties than we thought-so that reddishness is somehow included in the base…-it's not clear how subjectivity, the cognitive relation constitutive of a point of view, can be explained in terms of these properties. (2001, p.177)

In other words, subjectivity would remain equally inexplicable if the physical world included non-structural properties. The –tivity gap offers a plausible account of why this is the case: it is because the non-structural properties would still be *objective*, and there is an inevitable failure of entailment from the objective to the subjective.

We can defend the –tivity gap against the claim that it is the structural nature, rather than the objective nature, of the physical that makes subjectivity inexplicable. This is not deny that the structural nature of the physical presents an explanatory problem; it is just that this problem pertains to the *qualitative character* of phenomenal states and not to their *subjectivity*. Of course, all subjective states must have some qualitative character, therefore no subjective state can be fully explained in structural terms. Nevertheless, we must be clear about which *aspect* of phenomenal states is responsible for which explanatory impasse.

Parallel considerations arise in connection with the –trinsicality gap. The –trinsicality gap claims that the qualitative character of phenomenal states cannot be accounted for in physical terms. It also claims that it is the *structural* nature of physical states or, to put it another way, the *extrinsicality* of physical properties, that is responsible for this failure of entailment. Again, some defenders of the epistemic gap might affirm the former claim but deny the latter. Nagel (1974) holds that the *objectivity* of the physical makes an explanation of qualitative character impossible, and shows no concern for considerations surrounding the structural. This encourages an extension of the –tivity gap that incorporates the qualitative character of phenomenal states rather than just their subjectivity. Nagel argues as follows:

> …if the facts of experience - facts about what it is like for the experiencing organism - are accessible only from one point of view, then it is a mystery how the *true character* of experiences could be revealed in the physical operation of that organism. The latter is a domain of objective facts par excellence - the kind that can be observed and understood from many points of view and by individuals with differing perceptual systems. (1974, p.442, my italics)

In a sense, Nagel is right to hold that phenomenal character can never be explained in objective terms. Since phenomenal character belongs to subjective states, and subjective states cannot be explained in objective terms, phenomenal character cannot be explained in objective terms. However, Nagel's line of thought is potentially misleading in at least two ways.

First, Nagel's position implies that if the –tivity gap could be overcome, the qualitative character of phenomenal states would no longer be mysterious. This simply ignores the genuine gap between extrinsic physical properties and intrinsic phenomenal qualities. If we somehow found an explanation of why there is something it's like to occupy states with particular objective properties, we would be left with the mystery of how what it's like to occupy that state could be exhaustively determined by non-structural properties.

Second, to claim that the phenomenal qualities accessed from a specific point of view are inexplicable in objective terms implies that the existence of the point of view itself does not present an explanatory problem. Put another way, *what* it is like to be in that state is deemed mysterious, but there being *something* it's like to occupy a state is regarded as unmysterious. This is reflected in Nagel being comfortable asserting that there is *something* it's like to be a bat, but arguing that *what* it's like to be a bat is inaccessible to us (1974, p.438). This outlook distorts the real nature of the –tivity gap. Levine captures this: '…it seems to me that [Nagel] doesn't sufficiently appreciate that the entire idea of a point of view is itself deeply puzzling. A way to put the problem…is just this: how could anything like a point of view exist?' (2001, p.177) The fact that there is anything it's like at all to occupy a state is inexplicable in objective terms.[26] To capture why the objective nature of the physical generates an explanatory impasse, we should therefore refer to the *subjectivity* of phenomenal states as such rather than to the qualitative character of subjective awareness.

[26] There are further complications about Nagel's contribution to this discussion. It is not entirely clear what Nagel means by 'point of view' (Stoljar 2006, p.153). He does *not* seem specifically to have *phenomenal* points of view in mind. Furthermore, he is explicitly concerned with point of view *types* rather than *tokens* (1974, p.441). The *–tivity gap*, however, is concerned with the inexplicability of token phenomenal points of view in objective terms.

The –trinsicality gap can be defended against the claim that it is the objectivity of the physical, rather than its structural nature, that prohibits a physical explanation of qualitative character. This is consistent with the objectivity of the physical generating an explanatory impasse, but that impasse pertains to the existence of phenomenal states as such rather than to the character of those states, as captured by the –tivity gap. Phenomenal qualities may indeed be inexplicable in objective terms, but only because they are properties of subjective states, not because of some further complication that phenomenal qualities add to the situation.

The commitments of the two gaps and the relationship between them should now be more clear. These two gaps reflect existing insights into the epistemic gap, including Chalmers's work on the non-structural/structural divide and Nagel's work on the objective/subjective divide. However, my account of the two conceptual gaps diverges from existing positions in certain respects. I have by no means put forward a new account of why consciousness is inexplicable in physical terms, but I have argued for a shift in our understanding of existing key insights.

Other conceptual gaps between the physical and phenomenal have been proposed. McGinn (2004), for instance, draws on Descartes to argue that the non-spatiality of experience renders it inexplicable in spatial terms. It would take us too far afield to evaluate these proposals, but I will make the following general comment: these alternative conceptual gaps are either less compelling than the two we have established, or are merely variations on those gaps. As such, we can afford to put them aside.

3.3.2. The Ramifications of the Conceptual Gaps

The two conceptual considerations are meant to act as *vetoes* against the speculation that an improved epistemic position could close the epistemic gap.[27] Consequently, advocates of the epistemic gap can maintain that the gap is absolute whilst

[27] Judisch (2008, p.316) argues that a negative characterisation of the physical as non-phenomenal fails to capture the apparent gap between the physical and the phenomenal. It is worth noting that the *positive* characterisations of the physical as objective and structural allow Judisch's worries to be avoided.

acknowledging our less-than-ideal epistemic position. However, by appealing to the –tivity gap and the –trinsicality gap, the importance of CA and KA is diminished. Those arguments were supposed to reveal a failure of entailment from the physical to the phenomenal. The two conceptual gaps show that this failure of entailment is inevitable given what *kind* of fact physical facts are, and what *kind* of fact phenomenal facts are. But if these conceptual gaps reveal an inevitable failure of entailment, the original arguments are no longer required. Plugging the ontic conditional into the arguments TIV and TRIN would generate a conclusion of Primitivism independently. CA and KA might make the gaps more *vivid* but they are not where the real philosophical action is happening (see Stoljar 2006, p.155). In a sense, the conceptual gaps do not so much reinforce CA and KA as *replace* them.

Regarding CA, what *we* can conceive does not show that phenomenal properties are inexplicable in non-phenomenal terms. What an epistemically *ideal* subject can conceive might show this, but our insight into such a subject is provided by our appreciation of the conceptual gaps, not by any conceivability test that we perform ourselves. It is because those conceptual gaps reveal an inevitable failure of entailment that they put us in a position to draw conclusions about what an ideal subject can conceive. Similarly, KA relies on our having some grip on what Mary's complete knowledge is like (Alter 2009, p.761). We can characterise her knowledge as objective and/or as structural, but why should we believe that she is unable to deduce the phenomenal facts from such knowledge? It is only because we have an independent understanding of the conceptual gaps between the objective and the subjective, and between the structural and the non-structural, that we can be sure that Mary is unable to deduce the phenomenal facts. But if these conceptual gaps reveal the failure of entailment by themselves, it is not intuitions about Mary that justify a commitment to the epistemic gap.

Primitivists can, and must, defend themselves against the rudimentary response by appealing to the two conceptual gaps. But by making this move the dialectical situation undergoes an important shift. The task for the Physicalist is no longer to confront CA and KA face on. Instead, if they can cast doubt on the –tivity gap and –trinsicality gap, the path is open for the rudimentary response to explain away

the intuitions that drive CA and KA. Without the conceptual gaps, any sense that zombies and inverts are conceivable, and that Mary would learn something new, could be dismissed as a reflection of our limited knowledge of the physical explanation of consciousness. The case for Primitivism thus depends on the plausibility of the conceptual gaps.

Claiming that Primitivists must move from the 'initial' case to the 'refined' case by no means counts against the truth of Primitivism. Indeed, the two conceptual gaps have substantial *prima facie* value. Looking at the refined case for Primitivism, it is tempting to offer a straightforward answer to the question of consciousness: phenomenal properties *are* onticly independent of physical properties. However, we will soon see that matters are not so simple.

SECTION 4

THE CASE AGAINST PRIMITIVISM

The refined case for Primitivism concludes that the phenomenal is onticly distinct from the physical. This conclusion alone does not tell us a great deal about the place of consciousness in nature. How things are physically clearly influences how things are phenomenally, and vice versa. For instance, states of your brain can influence your conscious experience, and your conscious states can influence states of your brain and your bodily behaviour. The task for the Primitivist is to offer a metaphysical account of this apparent two-way interaction. Explaining these interactions in terms of phenomenal states *constituted* by physical states would only be an option for a Physicalist ontology. For the Primitivist, these interactions must be understood as causal interactions between distinct existences. Causal interactions between distinct existences are bound to laws of nature.[28] In order to make sense of physical-phenomenal causal interactions, the Primitivist should thus posit *psychophysical* laws

[28] I make no claims here as to *what* the specific relationship is between causation and laws of nature, though there will be some discussion of this issue in Chapter 4.

that govern such interactions (e.g. Chalmers 1996, p.213). I will put aside worries about the notion of psychophysical laws as such, and about whether there are plausible laws that could provide the requisite correlations between physical and phenomenal events. Instead, the issue I will focus on concerns phenomenal-to-physical causation. I argue that Primitivism is unable to provide a defensible account of these apparent causal interactions. By contrast, a Physicalist account is not affected by these concerns, which strongly encourages the rejection of Primitivism.

4.1. PHENOMENAL CAUSES AND PHYSICAL EFFECTS

Do conscious states have physical effects? If the Primitivist answers 'yes', then they are committed to a violation of the 'causal closure' of the physical, but such a violation is unacceptable. If the Primitivist answers 'no', then they are committed to 'epiphenomenalism', but that too is unacceptable. As such, phenomenal-to-physical causation raises a serious dilemma for Primitivism. I will consider each horn of this dilemma in turn before showing that a Physicalist stance is more plausible.

4.1.1. Efficacy and Causal Closure

When asked whether conscious states have physical effects, the *intuitive* answer would be 'yes'. From the first person perspective, it appears that our phenomenal states are part of the causal story behind at least some of our physical behaviour. Imagine looking at some paint and declaring "that paint is red, not purple". Intuitively, your having a perceptual experience characterised by red-qualities rather than by purple-qualities is a cause of this verbal behaviour. It seems, for instance, that if you had instead had a perceptual experience characterised by orange-qualities, you would not have behaved in that way. Furthermore, it seems that your being conscious *at all* is causally relevant to your actions. Putting aside the qualities of your experience, the very fact that you are conscious rather than unconscious seems to influence what you do. Our phenomenal states also seem to affect our non-verbal behaviour and our non-behavioural physiological states. In order to respect this, the Primitivist should

maintain that the phenomenal is physically efficacious: that phenomenal events can cause physical events to occur.

Primitivism claims that phenomenal events are non-physical – that they involve the instantiation of phenomenal properties that are onticly distinct from physical properties. Consequently, to hold that some physical events have *phenomenal* causes is to hold that some physical events have *non-physical* causes. However, the 'causal closure' of the physical is thought to prohibit physical events having non-physical causes. Causal closure says that all physical events have a complete physical cause.[29] When giving a causal explanation for a physical event, there will never be any need to step outside the physical domain.[30] We can apply this principle to the case of the utterance "that paint is red, not purple". The causal story here involves the paint reflecting light-waves onto the eye, information being transferred via the optic nerve to the brain, a sequence of neurological events in which that information is processed, and electrical signals being sent to the mouth and vocal chords that produce sound-waves we would recognise as the utterance "that paint is red, not purple". The details of the causal story are not important. What's important is the thought that there are no gaps at any point in the physical causal story. We have reason to believe each physical event finds a complete causal explanation in the event preceding it. There are no physical events that call out to be explained in terms of non-physical causes.

We can conclude that for any physical event with a putative non-physical cause, that physical event will have a complete physical explanation. This does not yet exclude the possibility of the physical event also having a non-physical cause, but it is easy to justify this further step.[31] If in all cases of phenomenal-to-physical causation, the physical event has a complete physical cause, that event will always be *overdetermined*.[32] This means that phenomenal states can only ever cause physical events that would have happened anyway. Can we really make sense of phenomenal

[29] See, for instance, Kim (1989, p.43) and Papineau (2002).

[30] This is a principle informed by physical theory. Though we have not defined the physical with reference to physical theory, it is appropriate to use physical theory as a guide to how things stand in the physical world i.e. in the *non-phenomenal* world (see Spurrett & Papineau 1999).

[31] In light of complications surrounding the metaphysics of causation, one might wish to put this point in terms of the *explanatory* irrelevance of the non-physical (e.g. Chalmers 1996, p.177).

[32] Kim (2002, p.177) argues that even the overdetermination model would violate causal closure because it is committed to possible worlds in which the physical cause is absent and the physical effect occurs with a complete non-physical cause.

states being physically efficacious if they only cause events that would still have occurred in their absence? It is doubtful that we can make sense of this kind of systematic overdetermination. As Kim points out (2002, p.174), if the physical event already has a complete physical cause, what causal work is there left for the phenomenal state to contribute? Furthermore, even if we could make sense of this, the causal role assigned to conscious states would be unsatisfactory. It does not merely appear to us that our experiences cause our physical behaviour, but that if we had not had that experience, we would not have behaved in that way. The overdetermination model cannot respect the second of these thoughts. Overall, the Primitivist cannot protect the physical efficacy of phenomenal states whilst respecting the causal closure of the physical. Rejecting causal closure, however, is not a viable option.[33]

4.1.2. Inefficacy and Epiphenomenalism

The Primitivist is now left with the other horn of the dilemma: denying that phenomenal states are physically efficacious. This 'epiphenomenalist' stance is deeply counter-intuitive. On this view, none of your physical behaviour is caused by your phenomenal states. Even the act of saying "I am phenomenally conscious" is not the result of your actually being phenomenally conscious. Epiphenomenalism also raises concerns about the *purpose* of phenomenal consciousness. If consciousness does not *do* anything, why do we have it? Given that our traits are the product of our evolutionary history, we should expect consciousness to have some survival value, but according to epiphenomenalism it can have no such value.[34]

These considerations, and many others like them, show that epiphenomenalism is an unattractive position, but not that it is false (see Chalmers 1996, p.160). We would need a compelling argument to justify the acceptance of such a strange position, but the combination of the case for Primitivism and considerations surrounding causal closure could be deemed to provide just such an argument.

[33] Overdetermination models are advocated by Mellor (1995) and Meixner (2004) but have been widely disregarded.

[34] A more general objection to epiphenomenalism is that '…science recognises no other cases of "causal danglers", ontologically independent states with causes but no effects.' (Papineau 2002, p.23)

Furthermore, the right Primitivist theory might be able to explain away the powerful intuition that our physical actions are caused by our phenomenal states. Particular psychophysical laws could guarantee an appropriate *correspondence* between our physical and phenomenal states even if the phenomenal states are ultimately inefficacious. For instance, the utterance "that paint is red, not purple" has a complete physical cause. Part of the physical explanation of that utterance is a non-phenomenal state of the speaker's brain that, thanks to some psychophysical law, also causes the occurrence of a conscious experience characterised by red-qualities (Chalmers 1996, p.159). A similar story will be told about the evolutionary origin of consciousness: the capacity for phenomenal consciousness does nothing useful, but could be lawfully correlated with physical capacities that *are* useful (Jackson 1982, Chalmers 1996, p.158).

It is unclear whether epiphenomenalism's defence against these initial objections can be maintained. Allowing that it can, there remains a deeper objection to epiphenomenalism that strikes at the heart of Primitivism. If phenomenal states do not have physical effects, problems arise concerning the possibility of *phenomenal knowledge*. Phenomenal knowledge is the knowledge we have of our conscious states. Of course, the case for Primitivism is founded on our phenomenal knowledge, so the epiphenomenalist cannot simply 'bite the bullet' and accept that we do not have phenomenal knowledge without fatally undermining their own position.

Here is a first pass at how epiphenomenalism precludes the possibility of phenomenal knowledge. Your zombie twin has no consciousness but is your duplicate in all physical respects. As such they have all the same functional states as you, including the state constitutive of believing that they are having a conscious experience. A cognitive story must be told about how your zombie twin comes to form this belief about themselves (see Chalmers 1996, pp.184-191). Whatever the cognitive mechanisms are that generate the belief, it is clear that they will not *justify* the belief. The zombie is victim of a deep cognitive malfunction when they form their false belief. The problem is that *your* phenomenal beliefs are the product of just the same cognitive processes as those of your zombie twin. Even if you are conscious, your phenomenal states cannot make any contribution to the belief-forming process. If your

zombie twin's phenomenal beliefs are not justified, nor are yours. As such, even if we *are* in fact conscious, we cannot *know* that we are. According to epiphenomenalism, we cannot then know that we aren't zombies.

This is a serious problem for epiphenomenalism. Considerations along these lines have even convinced Jackson to reject his own Knowledge Argument, since KA appears to undermine itself by precluding the possibility of Mary gaining phenomenal knowledge (Braddon-Mitchell & Jackson 1996, p.141).[35] However, Chalmers defends epiphenomenalism against this kind of objection by offering a more sophisticated model of how phenomenal knowledge works (1996/2003). The thought is that you and your zombie twin are *not* in the same epistemic position. There is something '…intrinsically epistemic about experience…', so by being conscious we have evidence for our phenomenal belief that is not available to our zombie twin (Chalmers 1996, p.196). Our phenomenal states can then *justify* our phenomenal belief without having any causal influence upon our functional states. The physical *portion* of you is no more justified in having the phenomenal belief than your zombie twin is, but as a conscious being you have both physical and non-physical components, and this composite subject of knowledge *does* have justification for their phenomenal belief.[36]

This defence of epiphenomenalism raises some important points. For instance, in Chapters 5 and 6 we will see that there are good reasons to believe that our knowledge of phenomenal states should not be understood on the model of other kinds of knowledge. Nevertheless, the account offered by Chalmers is unsatisfactory. We should concede that being in a phenomenal state gives you evidence for the belief 'I am conscious' that is not available to your zombie twin. However, it is one thing to be in possession of evidence that can justify a belief, and quite another to *use* that evidence in a way that justifies your belief.

[35] Nagasawa (2010) discusses this and concludes that the objection to KA should not be sustained.

[36] This position has deeply counter-intuitive metaphysical commitments: the physical world is such that we form phenomenal beliefs on the basis of flawed cognitive processes, but there happen to be non-physical phenomenal events that are nomically bound to our cognitive states in such a way that our beliefs generally come out true. Though this is a serious objection, it just adds to the stack of counter-intuitive implications of epiphenomenalism. By contrast, the epistemic objection under consideration promises to undermine the arguments that motivate such a position.

Consider a detective at the scene of a murder. The victim's body was found by a decorator who has been painting the victim's house. The decorator is actually the killer, but is playing innocent. However, the detective has a deep aversion to decorators, and immediately concludes that the decorator is the murderer. The detective inspects the knife in the victim's body, and notices that there is fresh paint on its handle. Unfortunately, the detective is not very bright, and fails to realise that this lends support to his belief that the decorator is the killer. Does the detective know that the decorator is the killer? It would seem not. The detective's belief is true, but the process that led him to form the belief fails to provide adequate justification. The detective is also in possession of evidence with the potential to justify his belief, but this does not help. It is not the case that the detective has the belief *because* of this good evidence, so the belief fails to qualify as knowledge.

Chalmers's account of phenomenal belief puts us in an epistemic position no better than that of the foolish detective. Our phenomenal belief is the product of the same misguided cognitive process as that of our zombie twin. Unlike our zombie twin, our belief is true and, thanks to the inherently epistemic nature of consciousness, we are aware of evidence that would justify our belief. But this is no better than the detective noticing the paint on the knife: if having good evidence does not play an appropriate role in the formation of a belief, it fails to justify that belief, and so fails to make that belief a case of knowledge.[37] Inefficacious phenomenal states might provide subjects with *evidence* for their phenomenal beliefs, but without the capacity to influence physical processes they cannot *justify* their phenomenal beliefs.[38]

We should not pretend that this objection to epiphenomenalism is straightforward. Bayne notes that '[t]he justification of phenomenal judgments raises some of the thorniest questions in epistemology…' (2011, p.418), and Chalmers has done significant work that attempts to rebut the kind of objection we have raised

[37] This is similar to Bayne's (2001, p.417) objection that the epistemic access a 'self' has to their own phenomenal states does not automatically justify the phenomenal beliefs that the subject forms.

[38] The view that phenomenal states are *constituents* of our phenomenal beliefs might be thought to give phenomenal states the requisite justificatory role (Chalmers 1996, p.204 and 2003). However, it is not clear that such a constitutive role would bestow the required justificatory status, and it is doubtful that all cases of phenomenal knowledge have phenomenal states as a constituent. Furthermore, there are worries about how the physical 'component' of a phenomenal belief could possibly be set up to meld with non-physical epiphenomenal events to form a composite belief state.

(2003). Nevertheless, the prospects for epiphenomenalism look very poor. As such, we can conclude that the second horn of the dilemma raised by phenomenal-to-physical causation cannot be taken.

4.2. FORMULATING THE PROBLEM

The Problem of Consciousness only arises if we have reason to prefer Physicalism to Primitivism. However, the fact that Primitivism is in trouble when it comes to the physical efficacy of phenomenal states does not, in and of itself, lend support to Physicalism. Why should we believe that Physicalism could provide the phenomenal with an appropriate causal status? Physicalism denies that phenomenal states are non-physical. As such, conscious experience falls within the causally closed system of the physical. Physical events, such as our bodily behaviour, have a complete physical cause. If phenomenal events in some sense *are* physical events, they can cause our behaviour (e.g. Levine 2001, p.5).

There are some complications with this picture. For instance, Kim (2002) suggests that even if mental states (such as phenomenal states) are constituted by lower-level physical states, it is those lower-level states that do all the causal work, meaning consciousness again becomes epiphenomenal. It is plausible, however, that arguments of this kind can be undermined given the right account of causation, or at least the right account of 'explanation' (Levine 2001, Yablo 2002). The key point is that on a Physicalist ontology there are not two causes – one physical and one phenomenal – competing to bring about an effect. Rather, there is one cause that can be described with a physical vocabulary but which can also be described in phenomenal terms. Overall, it is fair to conclude that causal considerations lend considerable support to a Physicalist view of consciousness.[39]

[39] Physical-to-phenomenal causation also generates serious difficulties for Primitivism. Psychophysical laws must account for why *our* physical states generate conscious experiences. This challenge generates the following dilemma: either the psychophysical laws are such that only physical states very close to our own can cause phenomenal events *or* our physical states do not have any special status in the psychophysical scheme of things. On the first route, psychophysical laws become unacceptably *ad hoc* and anthropocentric. On the second route, Primitivism ends up committed to 'panphenomenalism': the

As previously stated, a deep problem is revealed by the question 'is the phenomenal onticly dependent on the physical, or onticly independent of the physical?'. That problem can be captured as follows:

> ***The Problem of Consciousness:*** There are persuasive reasons to believe that the phenomenal is onticly independent of the physical, and persuasive reasons to believe that the phenomenal is onticly dependent on the physical.

CONCLUSION

We have now considered the key arguments that lead us towards these opposing verdicts on the ontic status of consciousness. It should be recognised that the case for Physicalism indicates that something must be wrong with the arguments for Primitivism, but does not reveal where they go wrong. So far the case for Primitivism is untouched. We started our discussion of consciousness, in Section 1.1, with a distinction between how consciousness *seems* from a first-person perspective and what consciousness *does*. It is the 'seeming' dimension that drives the case for Primitivism: our phenomenology appears to be inexplicable in physical terms. It is the 'doing' dimension that drives the case against Primitivism: the causal efficacy of consciousness appears to be inexplicable in non-physical terms. If we are to respect both dimensions of phenomenal consciousness, something must be done to resolve this antinomy.

---

view that phenomenal consciousness is ubiquitous. There are serious objections to each route. Furthermore, a parallel dilemma arises in connection to the influence our physical states have upon the *qualitative character* of our conscious states.

CHAPTER 2

RESPONSES TO THE PROBLEM

The purpose of this chapter is to evaluate the standard responses to the Problem of Consciousness established in the previous chapter. A proposed solution will either *defend* Primitivism from the objections raised against it, or attempt to undermine the case *for* Primitivism. For a variety of reasons, I will not consider attempts to defend Primitivism any further. First, the case against Primitivism already outlined is sufficiently strong. Second, in the debate between Primitivism and Physicalism, the burden of proof lies with the Primitivist. If nothing else, Occam's Razor puts the onus on Primitivists to show that a purely Physicalist ontology fails, and that basic phenomenal properties must be introduced into our ontology. Since the burden of proof is on Primitivism, if a defensible form of Physicalism can be provided then Primitivism should be rejected. The goal of this thesis is to develop a viable Physicalist account of consciousness. If the account is successful, that alone would thus provide us with sufficient reason to reject Primitivism. Third, extending the case against Primitivism will not significantly enhance our understanding of what a viable response to the Problem of Consciousness involves. By contrast, the evaluation of existing forms of Physicalism will yield important conclusions that will inform the overall trajectory of the thesis.

Strategies that attempt to undermine the case *for* Primitivism fall into two main categories. 'Type-A' positions reject the epistemic step of Primitivist arguments, denying that there is a genuine epistemic gap between the physical and the phenomenal.[1] I argue that the standard positions of this type either fail to address the problem at hand, or deny the manifest reality of phenomenal consciousness. 'Type-B' positions reject the ontic step of those arguments, denying that an epistemic gap between the physical and the phenomenal entails that they are onticly distinct. I argue that this category of response rests on a mistaken understanding of a posteriori

[1]The 'type' labels are taken from Chalmers (2002).

necessity.

I will conclude that standard responses to the Problem of Consciousness are unsatisfactory. Besides this negative conclusion, I will also establish three positive criteria that an adequate response to the problem should satisfy. I will state the first of these without further argument:

> ***The Physicalist Criterion:*** A defensible response to the Problem of Consciousness must not hold that the phenomenal is onticly distinct from the physical.

The second and third criteria will emerge in my evaluation of Type-A and Type-B positions respectively. The failure of the standard responses motivates the exploration of an alternative approach to the problem in Chapter 3 and the three criteria I establish will guide the evaluation of that strategy.

SECTION 1

TYPE-A RESPONSES

Type-A responses deny that there is a genuine epistemic gap between the physical and the phenomenal.[2] On this view, the psychophysical conditional 'P→Q' is an a priori truth (see Chapter 1, Section 2.1.1). Regarding CA, zombies and inverts are not genuinely conceivable. Regarding KA, Mary does not really learn anything new on having her first experience of redness. Furthermore, the –tivity and –trinsicality gaps fail to present any genuine conceptual chasm between the physical and the phenomenal. How can Type-A theorists overcome the wide-spread and compelling intuitions that there is a genuine epistemic gap? Asserting that *they* do not have the relevant intuitions or that, given the case against Primitivism, the intuitions *must* be in error, does nothing to diminish the force of those intuitions in others. The task for the Type-A theorist is to undermine the epistemic gap whilst acknowledging its *prima facie* force. They must provide *arguments* that cast doubt on the gap rather than appealing

[2]Advocates of Type-A positions might reject the ontic step of Primitivist arguments too (e.g. Dennett 1991a). If so, they are still most informatively identified as Type-A theorists rather than Type-B theorists.

to intuition.

We have already considered one such argument: the rudimentary response to Primitivism. By appealing to the potential of future science, and drawing analogies with apparent gaps that have ultimately been closed by science, the Type-A theorist can attempt to undermine the epistemic gap. However, we have also seen that the –tivity and –trinsicality gaps cast serious doubt on this response. Though the rudimentary response focuses on the physical side of the psychophysical conditional, the leading Type-A positions tend to focus on the phenomenal side. I divide these positions into two related kinds: Reductionism and Eliminativism. We will consider each in turn.

1.1. REDUCTIONISM

Reductionism aims to give an analysis of the concept 'consciousness' that makes consciousness amenable to explanation in physical terms. Analytic Functionalism, for instance, claims that consciousness is a functional concept: to be a conscious state is to have a certain functional profile, and the functional properties of that state determine the character of the conscious experience. Here consciousness is not analysed into physical terms directly, but the space is opened up for consciousness to be realised physically. There is no epistemic gap involved in the physical implementation of functional states, so if consciousness is a functional property, how things are physically can epistemically entail how things are phenomenally. That is, 'P→Q' would be knowable a priori.

The problem for Analytic Functionalism, and for other reductive analyses, is that they fail to do justice to our understanding of phenomenal consciousness. No functional (or related) analysis can capture the full nature of phenomenal consciousness. They inevitably leave an unexplained residue, and this phenomenal residue is responsible for the epistemic gap. To conceive of a zombie is to conceive of a being without that non-functional property that makes you *more* than a zombie. What Mary discovers is precisely what all her functional knowledge left out. The subjectivity of consciousness, and the intrinsicality of phenomenal qualities, refuse to be analysed

in functional terms. It seems that a reductive analysis of consciousness can only get off the ground by *ignoring* the very insights that underwrite the epistemic gap. Once you have fixed all the physico-functional details – all the details that might figure in a reductive analysis of consciousness – you will always be left with the further question of whether that physico-functional state is accompanied by phenomenal experience, and of what that experience is like for its subject (see Chalmers 1996, p.47). A reductive analysis is successful only if it closes off the possibility of such further questions.

One popular route for the Type-A theorist is to introduce a two-step analysis of consciousness. In particular, our concept of consciousness might be amenable to analysis in *representational* terms. After all, conscious states are plausibly a special kind of mental representation, and phenomenal character is plausibly bound to the content of conscious mental representations. Assuming such an analysis is plausible, how would this help the Type-A theorist? The hope is that representation can be explained in physical terms. Accounting for the intentionality of mental states may be a genuine philosophical puzzle, but many are optimistic about the tractability of the problem (see Jackson, quoted Davies 2008 p.27). Theories often account for representation in causal and/or 'teleosemantic' terms, which allow mental representations to be realised by physical states. If phenomenal consciousness is just a variety of mental representation, it too can be realised physically.

There is something of value in this line of thought, and I will explore Representationalist strategies further in Chapter 5. The basic appeal to representation, however, does not help the Type-A theorist. Collapsing the project of explaining consciousness into the more tractable project of explaining intentionality does not make the former project any easier – it just makes the latter project *harder*. If consciousness is representational, then an explanation of mental representation will have to face up to the conceivability of zombies: a being like you in all physical respects but devoid of phenomenal representations (Crane 2007, p.24). Similarly, if Mary discovers a representational truth on leaving her room, then some representational truths are non-physical (Alter, 2007). Furthermore, the –tivity and –trinsicality gaps will now count against the plausibility of a complete physical explanation of representation.

If consciousness is susceptible to analysis in representational terms, there is an epistemic gap between the physical and the representational. If, on the other hand, mental representation *can* be accounted for in physical terms, then consciousness must involve the instantiation of non-representational properties. Either way, the epistemic gap between the physical and the phenomenal remains untouched.

1.2. ELIMINATIVISM

Eliminativists deny the existence of phenomenal consciousness. They concede that conscious experience cannot be explained by any physical theory but, as Levine explains, '…that's not because of some lack in the theory. Rather, the problem is that conscious experience doesn't really exist.' (2001, p.128)[3] This kind of account takes the epistemic gap between the physical and the phenomenal seriously: the apparent residue left by physical accounts of consciousness is genuinely inexplicable in physical terms. The claim, however, is that this residue is a fiction. The apparent non-functional remainder does not exist. On the face of it, this is a bold and implausible position. The task for the Eliminativist is to give us reason to believe that phenomenal consciousness is an *illusion*. There are various ways in which this might be done, but they are all ultimately unpersuasive.[4]

One route for the Eliminativist is to regard phenomenal consciousness as a theoretical posit, then show that this posit is not warranted by the data. Following Dennett (1991a), one might focus on the data of verbal reports. People are inclined, for instance, to report that they have a special internal awareness characterised by ineffable qualities. A 'folk psychology' theory explains this verbal behaviour in terms of people *really having* such phenomenal states. Cognitive science, however, can offer a competing theory that explains such reports simply in terms of information-processing structures in the brain. For example, the disposition to make reports of a special 'direct

[3] Some who claim to be Reductionists only offer a reduction of consciousness in the non-phenomenal sense of the term. If they deny the existence of *phenomenal* consciousness, they are ultimately Eliminativists. There is thus an extent to which Reductionism and Eliminativism merge into one another (Chalmers 2002, p.251).

[4] Eliminativist positions include Dennett (1991a) and Rey (1997).

apprehension' could be accounted for in terms of a subject accessing information but having no contact with the mediating processes responsible for this access (see Chalmers 1996, p.172). The full cognitive account would be more complex than this, but the key point is that it promises to explain why subjects *think* they have phenomenal states without holding that they actually have them. Cognitive explanations along these lines would be more economical than an explanation that posits real phenomenal states.

The problem with this kind of Eliminativist argument is that it misrepresents the status of phenomenal consciousness. If the phenomenal was a theoretical posit, it would indeed come at too a high a price, and we would plausibly be better off without it (Levine 2001, p.133). However, the real reason for believing that phenomenal states exist is not anything like verbal data. Rather, it is the phenomenal states themselves. We know about phenomenal states and their qualities by *having* them, not by inferring their presence from some non-phenomenal data. Phenomenal consciousness is an immediate datum – perhaps our *most* immediate datum – and not a theoretical posit (Levine 2001, p.134). An Eliminativist might argue that this special immediate knowledge of consciousness is illusory – that we are really just zombies. However, there is plausibly no space in which to argue that consciousness is an illusion. Sometimes how things seem is not how they really are, but since consciousness *is itself the seeming* how could it transpire to be unreal? The epistemic status of experience renders any attempt to cast doubt on its existence futile.[5]

Eliminativists often denigrate such first-person data (e.g. Dennett 1991a, p.72). After all, science is driven by the principle that data must be open to public investigation. But this kind of response simply begs the question against those who claim to have phenomenal states that they know immediately from the first-person perspective (Chalmers 2002, p.251). They have no dialectical obligation to provide third-person evidence for the existence of first-person data. Conversely though, the Eliminativist is under no obligation to trust that such putative first-person data is genuine. Consequently, Eliminativists and their opponents must simply agree to

[5]Some of the most popular forms of Eliminativism appeal to representational notions (e.g. Rey 1997). However, as with the Reductionist appeal to representation, this approach just faces the old explanatory problems under a new guise rather than successfully overcoming them (Levine 2001).

disagree. This dialectical impasse need not count against the epistemic gap though. The arguments for Primitivism can be addressed only to those who 'take consciousness seriously' (Chalmers 1996, p.165). If you persuade yourself to deny the data of conscious experience, the case for Primitivism will have no grip on you. But if you acknowledge the reality of phenomenal consciousness, the appearance of an epistemic gap between the physical and the phenomenal is compelling, and the Primitivist has firm ground on which to build their case.

The failings of the standard Type-A positions point to a second criterion that a plausible response to the Problem of Consciousness must satisfy.

> ***The Phenomenal Realism Criterion:*** A defensible response to the Problem of Consciousness must acknowledge the manifest existence of phenomenal states.

Reductionism fails to satisfy this condition as it ignores the distinctive characteristics of consciousness that are responsible for the epistemic gap. Eliminativism fails to satisfy this condition as it explicitly denies the existence of phenomenal consciousness. Any serious attempt to undermine the epistemic gap must take the existence of phenomenal consciousness seriously.

SECTION 2

TYPE-B RESPONSES

Proponents of Type-B positions accept that there is an epistemic gap between the physical and the phenomenal. What they reject, however, is the inference from an epistemic gap to ontic distinctness. Type-B theorists hold that the psychophysical conditional is a necessary truth, but claim that it is an *a posteriori* necessity. On this view, the Primitivist arguments might show that 'P→Q' is not knowable a priori, but they fail to show that 'P→Q' is not necessary. In the first sub-section I introduce the notion of a posteriori necessity and explain how the Type-B theorist attempts to use this notion to undermine the case for Primitivism. In the second sub-section I argue that all a posteriori necessary truths must be knowable a priori for an appropriately

informed subject. As such, if there is a principled epistemic gap between properties then they must be onticly distinct. In the third sub-section I evaluate some responses to this argument against Type-B positions but maintain that they are unpersuasive. I conclude that Type-B positions fail to present a plausible case against Primitivism.

2.1. A POSTERIORI NECESSITY

The notion of a posteriori necessity was introduced by Kripke to account for the necessity of propositions such as 'water is H_2O'.[6] This proposition is true in all possible worlds, so is not contingent, yet cannot be established through conceptual analysis alone, so is not a priori. The discovery that water is H_2O is thus an a posteriori insight into a necessary entailment between the instantiation of H_2O and the instantiation of water. This discovery was made through empirical observation and inference to the best explanation – the hallmarks of science – rather than through conceptual analysis and logical derivation – the hallmarks, perhaps, of philosophy.[7] As Byrne explains, 'Kripke pointed out that the notions of necessity and a prioricity are distinct: the former is from metaphysics, the latter from epistemology…' (1999, p.372). The notion of a posteriori necessity thus opens the way for positing necessary truths without any epistemic commitment to the 'apriority' of that truth i.e. to its being knowable a priori.

Claiming that the psychophysical conditional is an a posteriori necessary truth has significant *prima facie* appeal. Our discussion of Type-A positions revealed that the pursuit of an a priori derivation from the physical to the phenomenal is wildly over-ambitious. A combination of taking the epistemic gap seriously and finding Primitivism implausible has led many philosophers to occupy the middle ground offered by Type-B positions. Within the debate surrounding the Problem of Consciousness, adopting some form of Type-B position is probably the majority stance (Davies 2008, p.38). I will

[6]The literature typically uses identity statements as examples of a posteriori necessary truths. An identity between the physical and the phenomenal would guarantee the necessity of the psychophysical conditional, but such an identity is by no means compulsory. As such, it is more appropriate to approach the discussion in terms of necessary truths in general, rather than specifically in terms of identity claims.
[7]Block and Stalnaker (2002) emphasise that conceptual analysis plays no role in our coming to know that 'water is H_2O'.

outline how Type-B positions seek to undermine each of the Primitivist arguments.

CA rests on the premise that if something is epistemically possible, then it is metaphysically possible. Our capacity to appropriately conceive of zombies and inverts reveals their metaphysical possibility, thus showing that phenomenal properties are onticly basic. Type-B theorists, however, challenge this premise. Zombies and inverts may well be epistemically possible, but they are not metaphysically possible. There is no possible world occupied by beings that are like you in all physico-functional respects but who differ from you phenomenally. Our concepts allow us to imagine scenarios that are metaphysically impossible.

Type-B theorists challenge KA in a similar way. They allow that Mary is ignorant of certain phenomenal truths before leaving her room, but claim that these phenomenal truths are nevertheless necessitated by the physical truths.[8] Type-B theorists often describe Mary's discovery as her learning an *old fact* in a *new way*.[9] The discovered phenomenal facts are just familiar physical facts accessed under a novel mode of presentation, rather than non-physical facts as the Primitivist would claim. What it's like to undergo a red experience is necessitated by the physical facts – the facts Mary already knew – but the a posteriori character of this necessitation means we should not expect Mary to be able to *deduce* what qualitative redness is like from the relevant physical truths. It is informative for Mary to learn, on leaving her room, that being in the neurological state 'N' of red-perception has this reddish qualitative character rather than that greenish qualitative character. However, it does not follow that the link between N and phenomenal redness is contingent: that there are possible worlds in which N is associated with phenomenal green-ness. That may be a *conceptual* possibility but, according to the Type-B theorist, it is not a metaphysical possibility. Mary, no less than us, can entertain metaphysically impossible scenarios.

Type-B theorists do not typically address the –tivity and –trinsicality gaps, but it

[8]Type-B theorists can also challenge KA by suggesting that the Mary scenario is metaphysically impossible. After all, according to Type-B positions our capacity to conceive of the scenario does not suffice to demonstrate its metaphysical possibility.

[9]For a useful overview of this response, see Stoljar & Nagasawa (2003). There is a variant of the 'old fact/new way' response that is concerned with Mary's acquisition of new concepts on leaving her room rather than on the a posteriori character of the psychophysical conditional. This kind of response to KA is ruled out by our use of the *experienced* Mary scenario, in which she already has a concept of phenomenal redness before leaving her room.

is fairly clear how they would respond to these conceptual gaps. They can concede that there is no a priori derivation from the objective to the subjective, nor from the structural to the non-structural. Nevertheless, they would maintain that the objective facts necessitate the subjective facts and that the structural facts necessitate the non-structural facts. The conceptual gaps rule out any a priori necessitation, but they do not thereby rule out necessitation *as such*. The Type-B theorist can hold that there are no possible worlds like ours in all objective and structural respects, but divergent from ours with respect to the phenomenal.

The Type-B view starts from standard cases of a posteriori necessity, such as 'water is H_2O', then claims that the psychophysical conditional is an analogous necessary truth. Any plausible version of the view must, however, acknowledge a certain degree of *disanalogy* between the standard cases and the psychophysical conditional. If the psychophysical conditional is like any other a posteriori necessity, why is there such a deep sense of mystery associated with the physical explanation of consciousness, and not with standard cases like the chemical composition of water? We are not worried about 'zombie H_2O worlds' which are like our world with respect to H_2O but in which water is not instantiated (Levine 2001 p.51). We are not faced with theorists who claim that the instantiation of water is something onticly over and above the instantiation of H_2O (Levine 2001, p.80). If the psychophysical conditional is a run-of-the-mill a posteriori necessity, we are owed an account of why it *appears* to be exceptional – of why we resist accepting it as a necessary truth.

There are various routes open to the Type-B theorist here. Perhaps we have a propensity towards dualist thinking that is innate, or an engrained cultural habit.[10] Perhaps our resistance has a deeper cognitive explanation: maybe the cognitive systems responsible for thinking about consciousness, and those responsible for thinking about the physical, are unable to mesh with one another. Where we can link together our information about water and H_2O in a satisfactory way, there is an inevitable cognitive dissonance involved in the thought that phenomenal states are nothing more than physical states. Without adjudicating on this debate, it is quite plausible that a defensible psychological account of our resistance could be provided.

[10]Papineau (2011, p.15) attributes the former view to Bloom and the latter to Rorty and Yablo.

With such an account, the psychophysical conditional could be regarded as a typical case of a posteriori necessity, associated with an atypical discomfort about that necessity. This discomfort alone is not enough to cast doubt on the necessity claim: discomfort is no grounds for metaphysics (Papineau 2011, p.19). Overall, Type-B positions show significant promise. I will argue, however, that they are ultimately unable to live up to that promise.

2.2. AGAINST BRUTE A POSTERIORI NECESSITY

I will argue that we should adopt the 'Apriority Thesis'. This is the thesis that if the psychophysical conditional is necessary, then it must be knowable a priori for an appropriately informed subject. To permit a posteriori necessities that are not knowable a priori, even under ideal epistemic circumstances, is to permit what I will call 'brute' a posteriori necessities. This general rejection of brute a posteriori necessities justifies the specific rejection of the Type-B theorist's claim that 'P→Q' is necessary but not knowable a priori.

2.2.1. The Functional Role Account

In order to get a grip on the thought that a posteriori necessities are ultimately knowable a priori, we should start with the 'functional role' account of a posteriori necessity advocated by Chalmers & Jackson (2001). They claim that the necessary truth 'water is H_2O' has a kind of context-dependent apriority. Knowledge of this truth comes in two stages. First, a subject needs a proper analysis of what water is. Chalmers and Jackson offer a *functional* account of the meaning of 'water' whereby water is whatever performs the *causal role* definitive of water. That role is being the x such that it is potable, transparent, quenches thirst, falls from the sky etc. Importantly, this part of the subject's knowledge is a priori. Second, the subject learns that the substance H_2O performs the water role in the actual world. This is the a posteriori insight that H_2O is the actual stuff that is potable, transparent etc. From this, the subject can then

deduce that water is H_2O.[11] We can summarise this water deduction (WD) as follows:

WD1) Water is whatever x plays the waterish role. (*a priori*)

WD2) H_2O is the x that plays the waterish role. (*a posteriori*)

WD3) Therefore, water is H_2O. (*follows with a priori necessity*)

This example provides a clear illustration of an a posteriori necessity that involves an a priori grasp of how meanings depend on our environment, plus an a posteriori knowledge of what our environment is in fact like, yielding an a priori deduction of the necessary truth. Since this knowledge involves the contingencies of how the world in fact is, it is quite unlike our a priori knowledge that all bachelors are male. It would be a mistake, however, to think that the a posteriori component of WD makes knowledge of its conclusion any less a priori in character. Once a subject has established the premises, they will find it inconceivable for water to be onticly distinct from H_2O. According to this model, there is no such thing as brute a posteriori necessity, only a priori necessity with a posteriori components.

There is some controversy surrounding the Chalmers-Jackson model of a posteriori necessity. First, it is not clear that our concept of water really has a functional analysis. McLaughlin, for instance, holds that it is 'very much an open question' whether such an analysis is available (2007, p.209). Second, it is disputable that conceptual analysis plays any role in our coming to learn a posteriori necessary truths (Block & Stalnaker 2002). Third, even if the Chalmers-Jackson account of 'water is H_2O' is accepted, we would need compelling reasons to conclude that *all* a posteriori necessity must be understood on this model, and that there are no cases of brute a posteriori necessary truths (Polger 2008, p.114).

It may well be that the Chalmers-Jackson model can be defended against these worries, but I suggest that a more secure case against the Type-B theorist can be made if we put the functional-role account of a posteriori necessity aside. The central tenet of the Chalmers-Jackson model is captured by the thesis that all necessary truths are knowable a priori for an appropriately informed subject. Call the thesis that if 'P→Q' is

[11] If there are any concerns about using an identity claim as the example, the 'is' of identity can be replaced with the more modest 'is nothing more than'.

necessary then it is knowable a priori the 'Apriority Thesis'. We can argue for the Apriority Thesis without attributing any special role to conceptual analyses, functional or otherwise.[12] The two key premises of the argument are 'The Redescription Requirement' and 'Semantic Rationalism' which I will outline in turn.

2.2.2. The Redescription Requirement

Take the Physicalist claim that all properties instantiated in the actual world W* are nothing over and above physical properties. We have already outlined what such an ontic claim amounts to in Chapter 1 (Section 1.3). Now we must consider an important commitment entailed by such a claim of ontic dependence. If physical properties exhaustively determine how W* is, this has consequences regarding sentences true at W*. The truth-value of a sentence depends on truth-makers, and truth-makers are properties instantiated in the world. Many sentences true at W* will involve exclusively physical terms. As discussed in Chapter 1 (Section 1.2), what qualifies a term as a *physical* term is contentious. Clearly terms from physical theory, such as 'mass' and 'electron', are physical terms. For our purposes, we can make do with the minimal commitment that all physical terms are *non-phenomenal* terms. Take 'P' to be the conjunction of truths expressible in a physical vocabulary. Clearly the truth-makers for P will be physical properties. A more interesting commitment of Physicalism is that sentences true at W* that are *not* phrased in a purely physical vocabulary must also have physical properties as their truth-makers. If the truth-makers of those sentences were non-physical properties, then Physicalism would be false. Kirk describes this commitment as follows:

> If there are any true statements about the world which are not expressed in terms of the austere vocabulary of P, then those statements are different ways of talking about—describing, explaining, and so on—exactly the same world as is specified by P. (2006, p.524)

Physicalism requires that any sentence true at W* that is not included in P must be a *redescription* of the states of affairs already specified by P. Kirk (2006) gives the useful

[12] I also avoid using 'two-dimensional modal semantics' in the argument. The technicalities surrounding two-dimensional semantics can, as Polger (2008) suggests, distract from the important philosophical issues. Also, two-dimensionalism has been used to defend Type-B positions (e.g. Block & Stalnaker, 2002) so is unlikely to provide a solid foundation for the Apriority Thesis.

example of the truths of microphysics and landscape truths. It is plausible that true descriptions of mountains are simply redescriptions of that which is already captured by a complete microphysics. This is why mountains are not ontological basics. If you included mountains in your fundamental ontology, you would be guilty of *double counting* – of adding entities to the world that had already been captured at a different level. Similarly, if Physicalism is true then the complete conjunction of phenomenal truths 'Q' must be a redescription of properties and states of affairs already captured by 'P'.[13] Physicalism, and any other claim of exhaustive ontic constitution, must respect this 'Redescription Requirement'.

2.2.3. Semantic Rationalism

The 'Redescription Requirement' reveals how an *ontological* claim about W*, such as Physicalism, has *semantic* commitments concerning all sentences true at W*. The next step in the case for the Apriority Thesis is to show how these *semantic* commitments have *epistemic* implications. Semantic Rationalism is the label Witmer gives to the view that grasping the meaning of a term puts you in a position to work out how that term contributes to the truth-conditions of sentences in which it appears (2006, p.202).[14] Supporters of the Apriority Thesis generally seem to adhere to something like Semantic Rationalism. Witmer explains an important implication of the position:

> If I fully understand terms T1 and T2, then I know a priori how to figure out whether or not they have the same semantic value. More precisely, for any given situation in which both T1 and T2 may be used in a sentence, I can, if given sufficient empirical information about the situation, determine from that information alone whether T1 picks out the same thing as does T2. (2006 p.203)

---

[13]A complication here is that it may not be possible to capture phenomenal truths by way of sentences. Our conception of phenomenal states plausibly exceeds any linguistic characterisation. For instance, phenomenal qualities are often claimed to be *ineffable*. To get round this concern, think of Q as a *mental* sentence or a complex thought. If that 'sentence' involves phenomenal concepts that do not correspond to any linguistic term, so be it. The important point is that Q is a truth-bearer, and that the properties instantiated in W* are its truth-makers.

[14]Semantic Rationalism, and related claims, are fiercely disputed by Type-B theorists. My aim is to show that it is plausible without attempting to engage in the various complex debates that surround it. This debate shades into Quinean arguments against the a priori/a posteriori distinction. Those who reject the distinction are not easily categorised as Type-A or Type-B theorists. Engaging with Type-Q ('Q' for Quine) positions will lead us too far astray (though see Rosenberg 2004, p.70 for cogent criticisms).

A subject who understands a pair of terms, and who knows the relevant non-semantic facts that contribute to the semantic value of those terms, is in a position to know whether the terms co-refer.[15] No further factors could contribute to whether the terms co-refer, so there is no space for their co-referentiality to be hidden to the subject. The inferential route to the conclusion that the terms co-refer may be very complex, so a subject with inadequate inferential skills would be unable to draw the inference. But for a subject who does have optimal cognitive abilities, the inference is available.

It is worth considering some examples. Understanding the term 'water' involves understanding how its reference depends on environmental factors. Combine this a priori understanding (which may or may not involve explicit conceptual analysis, functional or otherwise) with a full a posteriori knowledge of all relevant environmental factors, and you will be able to discern that 'water' refers to the same things as 'H_2O'. One can be a competent user of the terms 'Mark Twain' and 'Samuel Clemens' without knowing that they co-refer. However, full a posteriori knowledge of the relevant non-semantic facts would reveal that they co-refer. This includes the fact that there was a person named 'Samuel Clemens' by his parents who chose to publish his books under the pseudonym 'Mark Twain' (see Chalmers & Jackson 2001). Given enough information of this kind, one can infer that the two terms co-refer. Their co-reference is not some *further fact* that must be learned a posteriori. There are a wide-range of non-semantic factors that might contribute to the semantic value of a term: baptisms, causal connections, the history of a term, use of a term by the wider community, use of a term by experts. Understanding terms is compatible with substantial ignorance about such factors, but *if* a subject has full knowledge of those factors then they do not need any further information in order to discern whether their terms have the same referent.

Applying Semantic Rationalism on the level of sentences, to understand a sentence is to understand what it takes for that sentence to be true. A subject who understands two sentences S1 and S2, and who has complete relevant knowledge of the non-semantic facts, would be in a position to know whether the sentences are true

[15]This is related to the Fregean view that those with mastery of a term have a priori access to the sense of that term, which determines its reference.

in virtue of the same truth-makers. In other words, if S2 is a *redescription* of a state of affairs described by S1, a fully informed subject would be in a position to know this. The fact that S2 is a redescription of the state of affairs described in S1 is not something that an ideal subject would need to learn a posteriori. All the relevant information is contained in the knowledge they already have, so if S2 is a redescription the subject could establish this a priori.

2.2.4. An Argument For the Apriority Thesis

What does all this mean for the psychophysical conditional? In order to qualify as Physicalist, Type-B theorists must respect the Redescription Requirement. This means that in the psychophysical conditional 'P→Q', 'Q' must be a redescription of properties and states of affairs already specified by 'P'. The truth of 'Q' does not require the world to contain anything more than what 'P' requires. Now consider a subject who knows P (or at least knows the relevant sub-set of P) and who understands Q (or at least the relevant sub-set of Q). According to Semantic Rationalism, the subject should be able to establish a priori whether Q redescribes the states of affairs described by P. Non-semantic factors might contribute to what terms (or concepts) in Q refer to, but the subject's knowledge of P will include all those non-semantic factors.[16] Of course, if a subject can know a priori that Q is a redescription of the world described by P, then they can know a priori that 'P→Q'. The Argument for the Apriority Thesis (AAT) can be captured as follows:

> AAT1) If 'P→Q' is necessary then the truth-makers of 'P' exhaust the truth-makers of 'Q'.
>
> AAT2) An ideal subject with full knowledge of P, full understanding of 'Q', and optimal cognitive capacities, could determine a priori whether the truth-makers of 'P' exhaust the truth-makers of 'Q'.
>
> AAT3) If an ideal subject can determine a priori that the truth-makers of 'P' exhaust the truth-makers of 'Q', then the proposition 'P→Q' is

[16]Technically we must include some non-semantic factors that are not included in P. P does not include the fact that the physical facts included in P are *all* the physical facts. Furthermore, P does not include facts about indexicals. For the sake of this discussion, we can just add a 'that's all' clause, and all indexical facts, into the antecedent of the psychophysical conditional. They may not strictly be physical facts, but they clearly are not facts that require the instantiation of non-physical properties, so no compromise is being made on Physicalism (Chalmers & Jackson 2001).

knowable a priori for that subject.

AAT4) Therefore, if 'P→Q' is necessary then 'P→Q' is knowable a priori for an ideal subject.

The first premise is a simple way of capturing the Redescription Requirement on claims of exhaustive ontic constitution. The second premise is a straightforward application of Semantic Rationalism.[17] P will include all the environmental factors that contribute to the meaning of Q – the factors that one cannot uncover by conceptual analysis alone. To deny that all such factors are included in P is to concede that non-physical factors contribute to Q holding, which is not an option for the Type-B theorist.[18] The third premise should be uncontentious. If you know that what makes P true also makes Q true, then you can infer that if P is true then Q is true. The conclusion follows that 'P→Q' is knowable a priori for an appropriately informed subject. Claiming that the psychophysical conditional is an a posteriori necessity does not do what the Type-B theorist needs it to. If the epistemic gap between the physical and the phenomenal holds, then 'P→Q' is unknowable a priori, even for an ideal subject. But if 'P→Q' is unknowable a priori even for an ideal subject then, from the argument above, it is not necessary. Overall, if 'P→Q' is an a posteriori necessary truth, it is still ultimately incompatible with the epistemic gap. The inference from an epistemic gap to an ontic gap still stands.

The Apriority Thesis allows us to reject the Type-B theorist's challenge to Primitivist arguments. Regarding CA, conceivability and possibility must coincide for an ideal subject. If 'P→Q' is necessary, then 'P∧¬Q' is inconceivable for an ideal subject. Of course, what *we* can conceive provides only a limited insight into what an *ideal subject* can conceive. As such, the Primitivist must be able to argue that an ideal subject would still face the epistemic gap. This much has already been conceded in connection with the Rudimentary Response to Primitivism (Chapter 1, Section 3.3.2.), but now we can appreciate its importance. Regarding KA, Mary *is* an ideal subject, so she should be in a

[17]Note, there is no assumption that the ideal subject could provide any kind of *analysis* of their phenomenal terms/concepts here. It is *possession* of those concepts that is doing the real philosophical work. The argument is not committed to the view, advocated by Chalmers & Jackson, that if there is no conceptual analysis of 'consciousness' then there can be no a priori entailment from the physical to the phenomenal.

[18]Boutel (forthcoming) argues that regarding the reference-fixing facts as non-physical has surprisingly little effect on the debate. It would just require supplementing P with the non-physical reference-fixing facts, then asking whether their conjunction entails Q a priori.

position to know (some sub-set of) the psychophysical conditional a priori. But the Type-B theorist fails to cast doubt on the intuition that Mary learns something new – something that she could not determine a priori from her physical knowledge and her concept of phenomenal redness – on leaving her monochromatic prison.

The responses to the –tivity and –trinsicality gaps that I attributed to the Type-B theorist can also be rejected. If there is a necessary entailment from objective and structural truths to the phenomenal truths, then that entailment must be a priori for an appropriately informed subject. Our comprehension of the objective/subjective and structural/non-structural dichotomies indicates that no such entailment could be knowable a priori. In light of the Apriority Thesis, we can then conclude that no such entailment is *necessary*. It is these insights that justify the claim that an ideal subject would still find zombies and inverts conceivable, and that Mary would learn something new.

2.3. IS CONSCIOUSNESS AN EXCEPTION TO THE APRIORITY OF ENTAILMENT?

How might the Type-B theorist respond to the Argument for the Apriority Thesis? One option is to insist that a posteriori necessities are generally *brute* a posteriori necessities: that they are not generally knowable a priori for an ideal subject. Though there are some arguments for such a stance, engaging any further in that debate would be beyond the scope of this thesis. An alternative strategy for the Type-B theorist is to concede that the standard a posteriori necessities are knowable a priori, but to maintain that the psychophysical conditional is an exception to the rule. On this view, there is something distinctive about 'P→Q' that justifies the claim that it is a brute a posteriori necessity. This strategy comes in two forms. Necessitarian Dual Attribute Theory challenges the application of the Redescription Requirement to the psychophysical conditional. The Phenomenal Concept Strategy challenges the application of Semantic Rationalism.[19] Both positions will be rejected.

[19]The label 'Phenomenal Concept Strategy' is taken from Stoljar (2005).

2.3.1. Necessitarian Dual Attribute Theory

This is the view that '…phenomenal properties are distinct from, but strongly determined (necessitated) by, physical properties.' (Davies 2008, p.25) According to this proposal, 'P→Q' is a necessary truth, but it is *not* the case that 'Q' is a redescription of properties and states of affairs already captured by 'P'. The claim is that the truth-makers of Q are phenomenal properties – properties distinct from the truth-makers of P – but that physical properties necessitate their occurrence. There is a psychophysical 'law of metaphysics' that is true in all possible worlds, but which is not knowable a priori.[20] On this account, the first premise (AAT1) of the Argument for the Apriority Thesis is false. Without the premise that 'P→Q' is necessary only if 'Q' is a redescription of P, the conclusion that 'P→Q' is knowable a priori cannot be reached.

There are many reasons to be suspicious of this proposal.[21] First, if there are no minimal physical duplicates of W* that differ from W* phenomenally, in what sense are the phenomenal and physical properties instantiated in W* distinct? Plausibly, being distinct from the physical requires not being necessitated by the physical. Second, this position concedes a great deal to Primitivism. It may rule out the metaphysical possibility of zombies and inverts, but it comes with the problematic costs associated with Primitivism. If phenomenal properties are distinct from physical properties, the causal closure of the physical suggests that they cannot be physically efficacious. The fact that the phenomenal is necessitated by the physical, rather than bound to it contingently, is not enough to avoid the threat of epiphenomenalism (Davies 2008, p.25). Third, positing a distinctive metaphysical necessitation between the physical and the phenomenal is *ad hoc*. Why should we accept that such laws of metaphysics exist, and why should we accept that such a law holds between the physical and the phenomenal? The familiar Kripkean cases do not provide any precedent for such a relation, so good reasons would need to be provided for accepting that the psychophysical conditional has a special status. Furthermore, how could we come to know such a necessary entailment when we only have a posteriori knowledge of our

[20] I borrow the phrase 'laws of metaphysics' from Rosenberg's (2004, p.68) critique of the view.

[21] It is quite plausible that this position is 'emergentism' by another name. Serious objections to emergentism are easy to come by (see especially Kim, 1989) and could be turned against the Necessitarian Dual Attribute Theory.

own world, plus broadly conceptual knowledge, at our disposal? Overall, the implausibility of the Necessitarian Dual Attribute Theory serves to make AAT1 all the more plausible.

2.3.2. The Phenomenal Concept Strategy

This position has received a great deal of attention. It holds that the epistemic gap can be explained in terms of the distinctive cognitive status of our phenomenal concepts, and that there is no need to posit primitive phenomenal properties in order to account for the gap (e.g. Loar 1990). This position comes in many forms, but the most relevant variant seeks to qualify Semantic Rationalism, which is the second premise of the Argument for the Apriority Thesis.[22] The claim is that our phenomenal concepts – such as the concept of phenomenal redness – refer to physical properties. The psychological status of these concepts, however, means that no amount of physical knowledge would equip us to infer that our phenomenal and physical concepts co-refer. Many different accounts have been given of what this special psychological status is (for a useful review, see Chalmers 2007). We can illustrate the kind of account on offer with the example of Hill's (1997) view that phenomenal and physical concepts belong to distinct psychological faculties. The concepts are deployed in modes of reasoning that cannot appropriately interact with one another, so even where our phenomenal and physical concepts co-refer, we are unable to recognise this. As Davies explains, on this kind of view we have '…a duality of concepts without a duality of properties' (2008, p.3).

The strongest criticism of the Phenomenal Concept Strategy is offered by Chalmers (2007). In order to succeed, the strategy must account for our epistemic situation regarding consciousness, and to do so in purely physical terms. If the psychological story of our epistemic situation can be accounted for in physical terms, then a being like us in all physical respects should be in the same epistemic situation. Of course, our *zombie twins* are like us in all physical respects, but are they in the same epistemic situation as us regarding consciousness? Here the phenomenal concept

[22]One of the other variants rejects the general application of the apriority of necessity, and uses the cognitive status of phenomenal concepts to account for our resistance to the psychophysical conditional, as discussed in Section 2.1. of this chapter.

strategist faces a dilemma. On the one hand, they could deny that zombies are in our epistemic situation. After all, being conscious plausibly plays an integral role in our musings on the epistemic gap. However, conceding that our zombie twins do not have our epistemic situation means conceding that our epistemic situation does not have a purely physical explanation. But if the correct account of our epistemic situation involves non-physical properties, it will be of no use to the Type-B theorist. On the other hand, one could claim that zombies *are* in our epistemic situation. However, as we have already discussed (Section 1.2 of this chapter), our having phenomenal states contributes to our epistemic situation regarding consciousness. To leave consciousness out of the epistemic story is to leave out what makes the epistemic gap so intractable.

This objection is sufficient to cast serious doubt on the phenomenal concept strategy, but it is worth noting that the –tivity and –trinsicality gaps encourage a further objection. Assume that there is a genuine disconnect between our phenomenal and physical concepts, such that we cannot extract their co-referentiality even in ideal epistemic circumstances. This would explain why we are incapable of deducing the instantiation of phenomenal properties from the instantiation of physical properties. However, the –tivity and –trinsicality gaps do not merely make the *negative* claim that we are unable to perform such a deduction. Rather, they make the *positive* claim that physical properties are the *wrong kind* of property to be the referent of our phenomenal concepts. Even if we did suffer from the kind of cognitive blockage proposed by the phenomenal concept strategist, this would not undermine our insight that subjective states cannot be nothing over and above objective states, or that non-structural properties cannot be nothing over and above structural properties. These insights suggest that a being *without* the relevant cognitive blockage would still be faced with an epistemic gap.[23]

Overall, the Type-B theorist fails to undermine the case for Primitivism. The appeal to a posteriori necessity promised to block the inference from an epistemic gap to ontic distinctness. However, a proper understanding of a posteriori necessities suggests that such an inference is still sound. This conclusion allows us to put forward

[23]The spirit of this objection is captured in Rosenberg's comment (2004, p.41) that we have a positive insight into why consciousness cannot be physical rather than a mere cognitive blind-spot. Also see Levine (2001, p.84).

our third and final criterion for a satisfactory response to the Problem of Consciousness.

> ***The A Priori Entailment Criterion:*** A defensible response to the Problem of Consciousness must respect that Physicalism is true iff the psychophysical conditional is knowable a priori for an ideal subject.

CONCLUSION

This completes our survey of the standard responses to the Problem of Consciousness. The discussion of these positions has revealed three plausible criteria that a satisfactory response to the problem must satisfy. The positions discussed each fail to meet one or other of these criteria. In the next chapter I will consider a more promising kind of response.

CHAPTER 3

THE EPISTEMIC VIEW OF THE PROBLEM OF CONSCIOUSNESS

The aim of this chapter is to explore Stoljar's Epistemic View (EV) of the Problem of Consciousness. Though Stoljar's arguments will be our primary focus, I will also mention other positions that complement EV. In particular, points made by McGinn and Strawson are informative, though the positions they advocate diverge from Stoljar's (and from each other). EV claims that the phenomenal is not onticly distinct from the physical.[1] The appearance of a deep epistemic gap, it suggests, is symptomatic of our limited conception of the physical world. If we were equipped with certain physical concepts that we currently lack, it would be possible for us to account for consciousness in physical terms.

EV presents us with a novel strategy for confronting the Problem of Consciousness. Section 1 identifies EV's core claims and Section 2 argues that EV has significant *prima facie* promise. Section 3 concerns whether EV can live up to this promise and identifies the key obstacles to its successful implementation. I conclude that EV should be advocated iff it can satisfy two challenging conditions: the Relevance Condition and the Integration Condition. This lays the way for Chapter 4, in which I consider whether any version of EV is capable of satisfying these conditions.

SECTION 1

WHAT IS THE EPISTEMIC VIEW?

This section explores the fundamental idea behind the Epistemic View. The first sub-section identifies EV's core commitments. The second clarifies what kind of ignorance

[1] In order to be consistent with conclusions reached in Chapters 1 and 2, I will not always follow Stoljar's precise formulation of EV. Any divergences from Stoljar will be justified over the course of this chapter.

EV claims we are suffering from. The third explains how EV responds to the arguments for Primitivism.

1.1. THE IGNORANCE HYPOTHESIS

1.1.1. Ignorance and the Problem of Consciousness

The Epistemic View revolves around a hypothesis about our epistemic situation regarding consciousness: the ignorance hypothesis (Stoljar 2006, p.6). This is the hypothesis that we are deeply ignorant of a type of physical fact integral to the explanation of consciousness. The ignorance hypothesis is relevant to the Problem of Consciousness, EV claims, because if we are indeed in that epistemic position then the problem would be solved. The suggestion is that the case for Primitivism can be defused by citing such ignorance. Primitivist arguments are founded on the apparent epistemic gap between the physical and the phenomenal. EV claims that this appearance is merely a reflection of our impoverished epistemic situation, and that for an appropriately informed subject there is no epistemic gap. If EV succeeds in undermining the case for Primitivism, it breaks the antinomy constitutive of the Problem of Consciousness. Stoljar advocates a sharp distinction between the two claims at work in EV:

> …we may formulate the epistemic view as the conjunction of two theses. The first is a conditional thesis linking the ignorance hypothesis and the problem of experience:
>
> E1. If the ignorance hypothesis is true, the problem of experience is solved.
>
> The second is the categorical thesis about the antecedent of this conditional:
>
> E2. The ignorance hypothesis is true. (2006, p.6)

McGinn explains that if this kind of epistemic claim is true then, contrary to Primitivism, '[t]he world itself is as smoothly natural and seamless as one could wish; it is just that we lack the conceptual resources with which to discover its objective lineaments' (2004, p.64). This position recognises that physical states necessitating

phenomenal states appears mysterious to us, but McGinn explains that '…the sense of deep mystery we have, which naturally expresses itself in ontological rhetoric, is really entirely epistemic; the mystery is *relative* to the human intellect as it attempts to come to terms with the problem' (2004, p.64).

EV is perfectly compatible with our ignorance being relieved in the future.[2] We might acquire the 'missing concepts' and find ourselves in a position to account for consciousness in physical terms. Our failure to appreciate this possibility can be diagnosed in terms of our implicit belief that our conception of the physical world is quite comprehensive. Strawson explains that '…the idea that the mind-body problem is particularly perplexing flows from our unjustified and relatively modern faith that we have an adequate grasp of the fundamental nature of matter at some crucial general level of understanding…' (1994, p.105). It is this presupposition that EV seeks to challenge.

Does EV require our ignorance to be of the explanatory basis of consciousness, rather than of the phenomenal itself? According to Chomsky (2009), Priestley cited our ignorance of 'perceptions' in his response to the mind-body problem. However, a contemporary analog of this position would be implausible. Strawson explains that '…our acquaintance with the experiential simply doesn't leave room for us to make a mistake about its basic nature of such a fundamental kind that exposing the mistake could entirely dissolve the mind-body problem…' (1994, p.99). As such, EV is specifically a claim about our ignorance of the physical side of the psychophysical conditional.

Stoljar presents EV in terms of the unknown truths being 'non-experiential' (i.e. non-phenomenal) truths rather than *physical* truths. This is motivated in part by his conclusion that the notion of 'the physical' plays an inessential role in formulating the target problem. My conclusions in Chapter 1 (Section 1.2) echo some of Stoljar's concerns. There I argued for a *minimal* conception of the physical as non-phenomenal, so my use of 'physical' is not that different to Stoljar's use of 'non-experiential'. One reason, though, for preferring the term 'physical' is its connection with the causal

[2] EV is also compatible with McGinn's (1989) claim that our ignorance is permanent, which we will discuss further in due course.

closure of the physical. The case against Primitivism suggested that if consciousness involves non-physical properties, consciousness cannot be causally efficacious in the physical realm. If a proponent of EV claimed that the truths of which we are ignorant are non-physical truths, then they would likely be faced with the same epiphenomenalist commitments. In other words, if EV is defensible then the unknown truths had better be physical truths: that is, truths about non-phenomenal events within the causally closed system of the physical. This justifies putting Stoljar's preferences aside and formulating EV in terms of unknown *physical* truths.

1.1.2. The Explanatory Value of EV

EV seems pessimistic about how close we are to explaining consciousness, arguing that we do not even have the conceptual tools required to formulate such an explanation. However, EV is *optimistic* when it comes to solving the Problem of Consciousness, holding that the ignorance hypothesis solves the problem. Are these two claims compatible with one another? If we stay clear on what is required of a solution to the Problem of Consciousness, we should not be concerned that EV takes a negative view of the explanatory project (Stoljar 2006, pp.97-98).

The project of explaining consciousness is an *empirical* problem, on a par with projects such as explaining global warming (Stoljar 2006, p.42). These problems are *scientific* rather than philosophical, and they are solved precisely when a scientific explanation of the phenomenon in question has been provided. The *philosophical* problem, as explored in Chapter 1, is a very different kind of problem that requires a very different kind of solution. This problem is based on the fact that consciousness appears to be *inexplicable* - that it seems the empirical project cannot possibly succeed - and is concerned with the metaphysical implications of this explanatory impasse. The various responses to this problem can be seen as *philosophical* explanations of the *empirical* explanatory problem. In other words, they are *meta*-explanations.

Of course, one way of solving the philosophical problem would be to explain consciousness, thus overcoming the empirical impasse that seemed philosophically perplexing. This is not, however, a requirement. McGinn explains that:

> …to give a constructive solution would be to *produce* the property or theory that explains how the brain causes consciousness; but a non-constructive solution requires only that we find reason to suppose that such a property or theory *exists*, whether we can produce it or not. (2004, p.61)

As such, one can make a case for the *explicability* of consciousness without making a case for some specific *explanation*. The only thing the philosopher needs to explain is the appearance of inexplicability.

Primitivists claim that consciousness appears inexplicable in physical terms because it *really is* inexplicable in physical terms, but EV offers an alternative explanation. Stoljar argues that '…a hypothesis about our current epistemic situation is the best explanation for the distinctively philosophical predicament we are confronted with when we think about experience' (2006, p.10). EV attempts to explain away the apparent inexplicability of consciousness without attempting to explain consciousness. Our sense that there is an impassable epistemic gap merely reflects our own epistemic situation regarding consciousness, and does not reflect anything about its ontological status. If EV succeeds in explaining our philosophical predicament, we should no longer feel drawn to the *competing* views of the problem, such as Primitivism. Again, McGinn captures the situation succinctly:

> When we have the right explanation for our failure to solve the [empirical] problem we see why it is that the [other] options are not forced upon us, and thus we are relieved of the philosophical pressure they seem to exert. (2004, p.72)

It remains to be seen whether EV succeeds in this, but there is nothing wrong in principle with offering arguments in favour of the explicability of consciousness in physical terms whilst denying that such an explanation is available to subjects with our limited conceptual resources.

1.2. WHAT TYPE OF IGNORANCE?

1.2.1. Shallow Ignorance vs. Conceptual Ignorance

EV is obviously based on an epistemic claim, but it is important to be clear about what *kind* of epistemic claim it makes. Like most claims of ignorance, the ignorance

hypothesis asserts that there are propositions that we do not know. In this case, the propositions are those that describe the physical explanatory base of consciousness. There are two ways in which one might be ignorant of those propositions. First, one might be able to entertain the propositions in question, but be ignorant of their truth. Call this 'shallow ignorance'. Second, one might lack the concepts required to entertain the propositions in question. As with shallow ignorance, conceptual ignorance involves there being propositions that you do not know are true. Unlike shallow ignorance, if you are in a state of conceptual ignorance you cannot even *entertain* those truths. A consequence of conceptual ignorance is that it makes us ignorant of an entire *type* of truth: specifically, the set of truths which we could only represent if we had the relevant missing concept (Stoljar 2006, p.69). EV specifically claims that we are suffering from *conceptual* ignorance.

Stoljar clarifies this with reference to Russell: 'According to Russell, a blind person—that is, a person who by definition has not had the relevant experiences—is ignorant in a certain dramatic way about color' (2006, p.69). Our ignorance of the physical truths essential to the explanation of consciousness is claimed to be akin to this. EV holds that we have a conceptual blind-spot, and that the properties occupying this blind-spot are integral to the generation of phenomenal states. Asking us for the physical explanation of consciousness would be analogous to asking the blind person what colour grass is: neither we nor they have the concepts required to entertain the answer to the question.

It is worth noting that the blind person might be able to use colour terms with some competence and might have concepts that refer to colours. Nevertheless, there remains an obvious sense in which the subject has no colour concepts. To make sense of this, we can deploy Foster's (2008) distinction between *opaque* and *transparent* ways of knowing. Transparent knowledge tells you about the nature of a thing. Opaque knowledge tells you about a thing indirectly, but does not reveal its nature. The blind person may have an opaque knowledge of colour, but they do not know colour in a transparent way. EV hypothesises a failure of *transparent* knowledge. We may well have an indirect *opaque* knowledge of the hypothetical unknown physical properties,

but we have no concept of what those properties *are*. That is, we do not know those properties *transparently*.[3]

Conceptual ignorance is clearly much deeper than shallow ignorance. It is by making a claim of conceptual ignorance that EV attempts to improve upon the rudimentary response to Primitivism discussed in Chapter 1. The rudimentary response simply claimed that we do not yet have the correct theory of consciousness. The problem with this response was that it failed to do justice to the epistemic gap: a mere failure of explanation does not account for a compelling appearance of inexplicability. By contrast, EV's appeal to *conceptual* ignorance is supposed to provide a serious explanation of why consciousness appears inexplicable in physical terms. To understand how a claim of conceptual ignorance could account for the appearance of inexplicability, we should consider Stoljar's story of the slugs and the tiles.

1.2.2. The Story of the Slugs

Stoljar offers a useful analogy that sheds light on what the Epistemic View is proposing and why it is relevant to the Problem of Consciousness. The analogy is inspired by Jackson's story of the sea slugs. Jackson asks us to imagine that we discover a race of intelligent sea slugs at the bottom of the ocean. He explains that:

> Despite their intelligence, these sea slugs have only a very restricted conception of the world by comparison with ours, the explanation for this being the nature of their immediate environment. Nevertheless, they have developed sciences which work surprisingly well in these restricted terms. (1982, p.135)

The success of their science might lead some of these slugs to conclude that their science is in principle capable of explaining all natural phenomena, though from our perspective we know their conception of the world to be severely restricted. Jackson suggests we may be in a similar epistemic situation to the slugs, failing to realise that the explanation of consciousness is beyond our limited concepts.

Stoljar develops a related story that seeks to offer a more precise analogue of our proposed epistemic situation. He portrays a race of slugs that live on a mosaic

[3] I will return to this distinction in Chapter 4 (Section 2.3).

constructed from two sorts of tiles - triangles and 'pie-pieces' (i.e. segments of a circle). These tiles are configured to form a variety of shapes, but the slugs' perceptual access to them is limited to two shape-detecting systems. One detects triangles and the other detects circles. Stoljar argues:

> …given their access to the mosaic, it would be natural for these slugs to think that, at least so far as the tiles of the mosaic are concerned, it was constituted only by triangles and circles—of course this is a mistake, but it would be a natural one in the situation. (2006, p.1)

The slugs are *conceptually ignorant* of a certain *type* of truth: specifically, those that involve pie-piece tiles. Since it is truths of this type that explain the circle-truths, it would appear to the slugs that circles are *primitives* that are inexplicable in noncircular terms. Of course, this intuition would not be a reflection of metaphysical fact, but rather of their epistemic limitations. Perhaps the appearance that consciousness is onticly basic can similarly be explained in terms of our epistemic failings, rather than in terms of the world's actual ontic constitution.[4]

Stoljar extends the analogy with the idea of a slug monist who claims that circles must be reducible to triangles. This reductive project would inevitably fail, just like that of Physicalist Reductionism about consciousness (see Chapter 2, Section 1.1). The opponent of the slug monist would hold that circles are irreducible in principle, in much the same way as the Primitivist about consciousness does. Of course, both camps are mistaken. An account of the circular in noncircular terms *is* possible, but *not* in terms of triangles. Instead, the correct account would involve the pie-piece tiles, of which the slugs have no concept. Following the analogy through, perhaps an account of consciousness in physical terms is possible, though not within the parameters of our current conception of the physical world. We could only explain consciousness if we were to acquire some analogue of the pie-piece concept: some consciousness-relevant physical concept.

It is worth noting that the slugs' conceptual ignorance of pie-pieces only gets them into trouble because of the role the pie-pieces in fact play in the constitution of

[4] The slugs may well have an *opaque* conception of the pie-piece tiles. For instance, they have the concept 'tile' which includes pie-piece tiles in its extension. This modest kind of knowledge is quite consistent with them having no *transparent* knowledge of the pie-piece tiles.

circles. Being conceptually ignorant need not *always* generate philosophical puzzles analogous to the Problem of Consciousness. Rather, such ignorance only causes trouble when the unknown properties are appropriately integral to the explanation of the target phenomenon. As such, EV is not committed to predicting deep philosophical problems wherever we suffer from conceptual ignorance. Overall, the story of the slugs helps us clarify what kind of ignorance EV hypothesises, and how that ignorance could be relevant to solving the Problem of Consciousness.

1.2.3. Missing Concepts vs. Misconceptions

EV claims that our current conception of the physical is impoverished - that there are physical properties for which we are missing any concept. The claim that our current conception of the physical is incomplete should not be confused with the claim that we currently *misconceive* the physical in some way. There are at least two positions which hold that the apparent Problem of Consciousness is symptomatic of our *mis*conception of the physical. Neither position is convincing, and they should be carefully distinguished from EV.

The first position takes the Problem of Consciousness to make essential use of the notion 'physical', and claims that our conception of the physical is faulty. Chomsky argues that '[t]he mind-body problem can be posed sensibly only insofar as we have a definite conception of body. If we have no such definite or fixed conception, we cannot ask whether some phenomena fall beyond its range' (quoted Stoljar 2006, p.54). He suggests that since '[t]he Cartesians offered a fairly definite conception of body in terms of their contact mechanics…they could sensibly formulate the mind-body problem.' (quoted Stoljar 2006, p.54) Today, however, we have no such substantive conception. More recently, Chomsky has argued that '…the concept "physical facts" means nothing more than what the best current scientific theory postulates hence should be seen as a rhetorical device of clarification, adding no substantive content' (2009, p.199). This encourages a skeptical stance towards the so-called Problem of

Consciousness: on this view, the question of whether Physicalism about consciousness is true has no real content, so the problem cannot even get off the ground.[5]

How does the skeptical position differ from EV? EV does not claim that the concept 'physical' is a broken concept. Missing certain physical concepts does not entail a misconception of the physical any more than missing a concept of kumquat entails a misconception of fruit. The temptation to conflate EV and the skeptical stance is exacerbated by Chomsky reporting some sympathy for EV. He appears to present two different stances on the problem: '[t]here is no reason to believe that humans can solve every problem they pose or even that they can formulate the right questions...' (2009, p.185). EV advocates something close to the first stance – we cannot, with our current conceptual repertoire, solve the *empirical* problem of explaining consciousness. EV is not, however, committed to the more radical second approach according to which the explanatory project, and so the philosophical problem it generates, is somehow misconceived.

Why not put EV aside and adopt the skeptical position? The key failing of the skeptical stance is that it over-estimates the role that physical theory plays in formulating the Problem of Consciousness. As I argued in Chapter 1 (Section 1.2.), understanding 'the physical' in terms of physical theory is indeed problematic. However, to motivate the problem we need only conceive the physical as non-phenomenal and causally closed. As such, Chomsky's aforementioned analysis of 'physical facts' is either false, or irrelevant to the sense of 'physical facts' that drives the problem.[6]

There is a second possible position according to which the Problem of Consciousness reflects our misconception of the physical, though this position does not have any unequivocal exponents. It is the view that our concept of 'the physical' as such is not at fault, but our particular concepts of physical properties are problematic. The claim is that we need to *reconceptualise* the physical world in a way that allows the phenomenal to be accommodated. Nagel advocates something along these lines, arguing that '…we need new intellectual tools, and that it is precisely by reflection on

[5] Other advocates of this stance include Crane and Mellor (1990).
[6] Stoljar (2006, Chapter 3) rejects the skeptical stance in a slightly different way, though this is mainly due to a difference in terminology rather than any real divergence from the objection just presented.

what appears impossible…that we will be forced to create such tools' (1986, p.52).[7] Chomsky (2009) also presents arguments that complement this conclusion rather than the general skepticism discussed above. On this view, our current physical concepts distort the true nature of physical reality. Only when these concepts are *replaced* by better concepts will we be able to account for consciousness in physical terms.

This is not the kind of speculation with which EV should be concerned. Not only does EV allow that our concept of 'the physical' is sound, it allows that all our current physical concepts are in good working order. EV simply claims that we could only explain consciousness if these physical concepts were *supplemented* by new physical concepts that filled-in our current blind-spots. Acquiring the relevant missing concepts need not alter our current physical concepts any more than acquiring a concept of kumquat would alter your concept of apples.

The reason we should not put EV aside in favour of the reconceptualisation approach is that it is not clear that there is anything wrong with our existing physical concepts. Why should we believe that when we use concepts like 'electron' and 'mass' we misconceive the physical world? The fact that we cannot explain consciousness using such concepts does not give us reason to believe that they misrepresent the world. Furthermore, why believe that any reconceptualisation could open the way for a physical explanation of consciousness? It is implausible that there could be concepts that achieve everything our current physical concepts achieve *and* somehow allow us to account for subjective qualitative awareness. This is not the kind of speculation we should take seriously, and has little relevance to EV. The ignorance hypothesis is concerned with what we know about the physical world rather than how we think about it: it claims that there are physical properties of which we are conceptually ignorant, not that we need to think about familiar physical properties in a new way.

1.2.4. Basic vs. Intermediate Ignorance

There is an intuitive idea that the physical world can be divided into levels. Physics, for instance, is thought to deal with the bottom level while astronomy deals with a much

[7] Nagel (1986) also argues that a reconceptualisation of the phenomenal is required, though McGinn (1989) offers convincing objections to Nagel's reflections on this.

higher level. Drawing on this notion, Stoljar makes the following distinction:

> Basic-level ignorance is ignorance of a truth concerning a particular element of nature or a basic fact of the world. Intermediate-level ignorance is ignorance of a truth that is not itself basic but is determined by the basic facts. (2006, p.72)

Stoljar claims that EV is neutral on whether we suffer from basic-level ignorance or intermediate-level ignorance.[8] I claim that we should take EV to be committed to our suffering from basic-level ignorance. Bennett (2009) argues persuasively that EV should not rest on a claim of intermediate-level ignorance. Take it that the C-truths are the basic truths, the A-truths are of a higher level and the B-truths mediate between them. Bennett asks us to imagine a subject 'Iggy' who knows the C-truths and wrongly believes that the C-truths do not entail the A-truths (2009, p.768). If his ignorance of B-truths is responsible for his error, we are left with the further question of '…why he mistakenly thinks it is possible for the *B*-truths to vary independently of the C-truths.' (2009, p.769)

If the explanation of Iggy's error is that he is ignorant of a particular type of C-truth relevant to the B-truths, then his original mistake is ultimately down to basic-level ignorance. If Iggy is ignorant of some fundamental principle that governs how C-entities combine to form B-entities, this would again be a case of basic-level ignorance since the principle itself is basic (Bennett 2009, p.768). If the error is to do with Iggy not thinking things through adequately, or to do with the brute a posteriori character of the entailment from C-states to B-states, this would just be a reiteration of the standard Type-A and Type-B positions we have already rejected. Overall, it is not clear that intermediate-level ignorance could explain modal errors without collapsing into a claim of basic-level ignorance or into something akin to one of EV's failed competitors.[9]

---

[8] Though the story of the slugs is a case of basic-level ignorance, Stoljar provides an illustration of intermediate-level ignorance involving intelligent moths (2006, pp.73-74).

[9] As Bennett (2009, pp.769-770) and Stoljar (2009, p.781) observe, this issue for intermediate-level ignorance is related to the debate surrounding the apriority of the psychophysical conditional. We will see in Section 1.3.2 that Stoljar's take on apriority is different to ours, which might explain his openness to a hypothesis of intermediate-level ignorance.

1.3. EV AND THE ARGUMENTS FOR PRIMITIVISM

So far, we have considered what EV consists in and outlined how the ignorance hypothesis could cast doubt on the apparent inexplicability of consciousness in physical terms. It is worth spelling out how EV addresses the arguments for Primitivism as outlined in Chapter 1. This should make it clear why one would claim that if the ignorance hypothesis is true, then the Problem of Consciousness is solved. Nothing so far is intended to justify EV's further claim that the ignorance hypothesis is in fact true. That will be considered in due course.

1.3.1. EV's General Response to Primitivism

As discussed in Chapter 1 (Section 2.1), arguments for Primitivism involve two key stages: the epistemic step and the ontic step. EV casts doubt on the epistemic step. It claims that the apparent epistemic gap between the physical and the phenomenal is an illusion symptomatic of our limited conception of the physical. This respects the ontological conditional that *if* there is an epistemic gap *then* there is an ontic gap, but it rejects the antecedent of that conditional. EV is thus a Type-A position.[10] In our initial discussion of the Type-A approach, we saw that the Type-A theorist must be able to explain why there *appears* to be an impassable epistemic gap between the physical and the phenomenal. EV has an intriguing response to this challenge.

EV can respect the fact that there is an epistemic gap between the-physical-*as*-*we*-*conceive*-*it* and the phenomenal. It is true that the types of fact that can be captured by our current conception of the physical are not the types of fact that could possibly entail the phenomenal facts. What EV denies is that there is an epistemic gap between the-physical-*as*-*such* and the phenomenal. If we had the hypothetical missing concepts – if all the types of physical fact relevant to the explanation of consciousness were available to us – there would no longer appear to be an epistemic gap. The

[10] EV could also be characterised as a Type-C position. According to Type-C positions '…there is a deep epistemic gap between the physical and phenomenal domains, but it is closable in principle.' (Chalmers 2002, p.257). However, Chalmers goes on to explain that Type-C positions generally '…collapse into one of the other views on the table…' (2002, p.258). Since it is clear that EV is ultimately a kind of Type-A position, the 'Type-C' label is superfluous.

reason we have the intuition that the physical-as-such cannot be responsible for consciousness is our implicit belief that there is no difference between the physical-as-we-conceive-it and the physical-as-such. We fail to acknowledge the possibility that we have a limited conception of the physical. We take it is a given that there are no types of physical fact beyond our current conception, or at least that we have enough of a grip on what any new type of fact would be like that we can be sure they are irrelevant to consciousness. Of course, according to EV this assumption is mistaken. We are like the slugs who mistakenly assume that the noncircular tiles to which they have access are the *only* kind of noncircular tile.

We previously noted that the rudimentary response to Primitivism can be countered by arguing that *more of the same won't do* – that more of the same kind of physical discovery will inevitably leave the epistemic gap untouched. Again, EV can respect this insight since further discoveries that deploy only our current physical concepts will inevitably yield the wrong kind of physical fact to close the epistemic gap. However, according to EV more of something *different* might allow the gap to be closed. The unknown physical truths relevant to EV are precisely *not* the kind of truth with which we are familiar. We only believe that no future discoveries could close the gap because we have a limited grasp on the kind of physical truths that such discoveries could disclose.

Of course, the arguments for Primitivism do not just *state* that there is an epistemic gap. They deploy various considerations designed to reveal that gap. Whatever consideration is deployed, EV's response is the same: for a subject with a relevantly complete conception of the physical, the consideration would have no force. As such, those considerations only have a grip on us because we erroneously take our limited conception of the physical to be relevantly complete.

1.3.2. Stoljar on EV and A Priori Entailment

It is worth noting that Stoljar has a different take on the status of EV. He recommends not taking a stance on whether the psychophysical conditional is an a priori truth or an a posteriori truth (2006, p.197). On this view, EV is not specifically directed at

challenging the epistemic gap. Stoljar claims that undermining our resistance to the psychophysical conditional has little to do with whether it is a priori or a posteriori, and that EV is the best way of undermining our resistance either way. As such, EV is neutral between Type-A and Type-B views. We have seen how EV could lend support to a Type-A view, but how could it help the Type-B view? Stoljar notes that if the psychophysical conditional is an a posteriori necessity, this alone would not explain why the psychophysical conditional appears to be contingent (2006, p.196). Additional claims are required to undermine the appearance of contingency (as we discussed in Chapter 2). Stoljar claims that standard attempts to provide the requisite additions fail, and that the ignorance hypothesis is the best way of explaining our mistaken belief that the psychophysical conditional is not necessary.

Though Stoljar's arguments raise some important points, I think it is appropriate to regard EV as a Type-A position. To justify this conclusion, we must address some of Stoljar's comments. First, Stoljar claims that the debate over (what we are calling) the Apriority Thesis is irrelevant to the plausibility of EV. It is not clear that this is true. If proponents of EV allow the possibility that the psychophysical conditional is a posteriori, this would open the flood gates to a wide range of Type-B positions to compete with EV. Stoljar argues that such positions fail (2006, Ch.9). However, it is hard to provide conclusive objections to these positions on a one-to-one basis. The Apriority Thesis provides the all-purpose objection that *no* version of Type-B position is satisfactory, and without this EV will have difficulty ruling out all the other Type-B proposals.[11] Dialectically speaking, EV is thus better off respecting the Apriority Thesis and identifying itself with the Type-A camp.

Second, Stoljar claims that the arguments for the Apriority Thesis are inconclusive (2006, p.197). We can agree with Stoljar that the issues here are especially wide-reaching and controversial. Nevertheless, we have an argument in favour of the Apriority Thesis, and Stoljar does not offer any explicit objections to that kind of argument. Given our commitment to the A Priori Entailment Criterion for a satisfactory response to the Problem of Consciousness, we are justified in considering only those versions of EV that promise to satisfy that criterion. Even if EV does have

[11] An objection along these lines is persuasively made by Papineau (2007).

the potential to remain neutral between Type-A and Type-B positions, we have independent reason to prefer a Type-A version of EV.[12]

Third, Stoljar argues that EV must be sharply distinguished from standard Type-A positions, and that those standard positions are mistaken (2006, Chapter 10). Here we can simply agree with Stoljar. In Chapter 2 (Section 1) we saw how traditional Type-A positions fail, and in this chapter we have seen that EV promotes a quite different strategy. However, this conclusion is compatible with understanding EV as a distinctive form of Type-A theory. EV – or at least the best version of EV – denies that there is an epistemic gap. This qualifies EV as a Type-A theory regardless of how different it is to the other positions in the Type-A camp. Overall, we should evaluate EV as an attempt to undermine the epistemic step in the arguments for Primitivism.

1.3.3. EV and the Conceivability Argument

Can we really conceive of zombies and inverts? EV suggests we cannot. We do not have the conceptual resources required to imagine a being like us in all physical respects, or even to imagine a being like us in all those physical respects relevant to consciousness. As such, we cannot fully conceive of a being like us in all physical respects but which differs from us phenomenally. Why, then, do we think that we can conceive of zombies and inverts? In our original exposition of conceivability tests (Chapter 1, Section 2.2.1) we identified two of the standard ways in which conceivability tests might misfire: proposition confusion and mode confusion. EV makes it plausible that we are making one or other (or both) of these mistakes when we claim to conceive of zombies or inverts (Stoljar 2006, pp.80-82).

Proposition confusion is where you conceive of some possible scenario *p* but believe you are conceiving some other scenario *q*. If *p* is possible and *q* is impossible, this will lead you into a modal error. According to EV, we may well be able to conceive of beings like us in all those physical respects captured by our current conception of the physical, but who differ from us phenomenally. After all, EV accepts that the physical-as-we-conceive-it cannot entail phenomenal states. The mistake comes with

[12] Interestingly, in Stoljar's earlier (2001) formulation of EV he treats the Apriority Thesis as a given, and claims that respecting that thesis is an important virtue of his position.

our belief that such an imaginative act constitutes conceiving of a being like as in *all* physical respects but different from us phenomenally. We confuse imagining *partial* physical replicas with imagining *complete* physical replicas. EV suggests that if we *really did* imagine a complete physical replica of ourselves – if we had the conceptual resources to factor in all the physical properties relevant to consciousness – then we would find it inconceivable for that replica to differ from us phenomenally.

Mode confusion occurs when you believe that you are *strongly* conceiving of *p* when in fact you are only weakly conceiving of *p*. Since weak conceivability does not entail possibility, this confusion can lead to modal error. Strong conceivability demands a substantive grasp on the scenario you imagine but, according to EV, we do not possess the conceptual resources to have such a grasp on the zombie or invert scenarios. It may well be that we entertain those scenarios and no contradiction occurs to us, but this would merely be a case of *weak* conceivability. The belief that this superficial reflection is a case of *strong* conceivability is what leads us into error. EV claims that if we really were strongly conceiving of a physical replica, we would find it inconceivable for them to differ from us phenomenally.

Stoljar points out that it is quite plausible that if EV is true then *both* kinds of error are at work when we take ourselves to conceive of zombies or inverts (2006, p.82). It need not be just one or the other. By citing these standard errors, EV avoids an implausible general skepticism about using conceivability tests as a guide to metaphysical possibility. Instead, EV makes it plausible that CA misfires in a way that conceivability tests often misfire. As McGinn explains, '[o]ur modal faculty naturally goes haywire in the conceptual vacuum generated by our ignorance' (2004, p.51).

Stoljar illustrates his proposal with the story of the slugs (2006, p.81). It is plausible that the ignorant slugs would claim to be able to conceive a mosaic identical to the actual mosaic in all noncircular respects but which lacks circles. However, it is clear that for a slug with a concept of the pie-piece tiles this scenario would be inconceivable. The ignorant slugs could be misled by proposition confusion: they confuse conceiving of a mosaic like the actual mosaic in all *triangular* respects with conceiving of a mosaic like the actual mosaic in all *noncircular* respects. Furthermore,

where the slugs claim to be able to strongly conceive of the mosaic like the actual mosaic in all noncircular respects, they only really have the conceptual resources to weakly conceive that scenario.

1.3.4. EV and the Knowledge Argument

EV claims that when we reflect on the Mary scenario, we are guilty of the same kind of mistake as we are with CA (Stoljar 2006, pp.82-83). Our intuitions about the Mary scenario would be askew if we had no grasp of a type of physical truth integral to the explanation of consciousness. We imagine Mary acquiring a wealth of physical knowledge far beyond our own but, plausibly, we imagine that the truths she learns are the same *type* of physical truth that we know i.e. that they do not involve the instantiation of any physical properties beyond our current conception. The pseudo-Mary we imagine would indeed be unable to infer the phenomenal truths from the physical truths. But we are not in a position to conclude that Mary-*proper* would be in the same situation. According to EV, Mary-proper would have knowledge of a type of physical truth beyond our current conception, and she would not learn anything new on leaving her room.

Here our mistake could be in taking ourselves to have a firm insight into Mary-proper when really we only have a firm insight into pseudo-Mary. This would be a form of proposition confusion. Alternatively, we might take ourselves to have a firm insight into Mary-proper when really we have only a partial insight into what her complete physical knowledge would involve. This mistake would be analogous to mode confusion. Stoljar notes that it is a virtue of EV that here it '…provides a unified treatment of CA and KA' (2006, p.82).

Again, the slugs shed light on things. It would plausibly appear to the slugs that it is possible for a 'superslug' to know all the noncircular truths and yet come to learn the circular truths separately. When reflecting on the superslug, they might imagine a slug with complete knowledge of the triangle truths, and they would be right in believing that such a slug would be unable to establish the circle truths a priori. But this would not be to imagine a genuine superslug. The genuine superslug would have a

concept of pie-piece tiles, and would be able to infer the circle truths from the noncircular truths. If a slug insists that this is not the case, it is only because they take themselves to have a better grasp than they really do on the epistemic situation of the superslug.

Our account of the case for Primitivism in Chapter 1 went beyond CA and KA. There we concluded that these two standard arguments should be supplemented by an appeal to the –tivity and –trinsicality gaps. In order to undermine the case for Primitivism, EV must therefore address those two conceptual gaps. The options for EV here are clear. Regarding the –tivity gap, the first option is to claim that there are unknown types of objective truth integral to the physical explanation of the subjectivity of conscious states. The second option is to claim that the unknown physical truths are subjective, so the gap between the objective and the subjective does not apply to the psychophysical conditional. Regarding the –trinsicality gap, parallel options are available. Either the unknown properties are extrinsic properties relevant to explaining the intrinsic properties of conscious states, or they are intrinsic properties and the –trinsicality gap no longer applies. I will postpone full discussion of this important issue until Section 3.2.

SECTION 2

WHY IS THE EPISTEMIC VIEW WORTHY OF ATTENTION?

Before we evaluate whether EV can be implemented in a satisfactory way, there is more to say about why it is an attractive strategy. In this section I will consider the key virtues of EV. This discussion will also serve to clarify EV's relationship to existing responses to the Problem of Consciousness.

2.1. THE THREE CRITERIA OF SUCCESS

In Chapter 2 I argued for three criteria that a satisfactory response to the Problem of Consciousness must meet: the Physicalist Criterion, the Phenomenal Realism Criterion

and the A Priori Entailment Criterion. The attractiveness of EV lies in its promise of satisfying all three of these criteria. EV protects the psychophysical conditional, and so advocates a form of Physicalism. It promises to achieve this whilst leaving our understanding of consciousness untouched, meaning there is no risk of it denying the existence of phenomenal states. Finally, it is consistent with the psychophysical conditional being a priori for an appropriately informed subject. This suggests that EV has a significant potential advantage over existing attempts to respond to the problem, and deserves to be taken seriously.

Primitivism, standard Type-A positions and Type-B positions are each capable of satisfying two of the three criteria, but only at the expense of the remaining condition. This puts EV in a strong dialectical position. The three criteria are plausible conditions and, all things being equal, we should prefer a response that allows us to accommodate all of those conditions over any response that asks us to reject one or other of them.[13] It is worth noting that this advantage does not presuppose that the three conditions are completely beyond question: the claim is simply that they are highly plausible, and that a sound methodology would suggest we only compromise one of them once the possibility of respecting all three has been ruled out. If EV cannot be ruled out, we have reason to prefer it over the more radical positions on the table.

In light of its promise to respect all three criteria, EV has the potential to talk across major divisions in the debate surrounding consciousness. It is one thing for a response to seem plausible to a disinterested party, but quite another for it to persuade those who are entrenched in one or other of the pre-existing categories of response. Hardcastle, in her review of Stoljar, argues that we should judge a proposal by '…how well it would convince those on the other side of the ideological divide…', advising that '[o]ne must speak across the divide, not sing platitudes to the choir' (2008, p.274). She then goes on to make a somewhat uncharitable case *against* EV's success in this respect, but on closer consideration it appears that EV shows great promise here.

[13] In his earlier formulation of EV, Stoljar argues that this 'accommodationism' is an important virtue of his position, though the criteria he claims to accommodate are slightly different to ours (2001, pp.278-9).

An initial benefit of EV is that it does not ask advocates of the competing positions to compromise on the key principles that motivate those positions. Roughly speaking, advocates of the standard positions are happy to sacrifice one of the three criteria of success because they believe that the other two criteria must be protected, and that protecting them requires rejecting the remaining criterion. Whichever of the criteria an existing position respects, EV can respect them too but without having to reject any of them. As such, if advocates of the competing positions defect to EV, it is likely they can take their driving philosophical principles with them.

A further benefit of EV is that, though it diverges from the standard responses to the problem, it displays something of the spirit of each of them. Again, this helps EV to talk across some deep philosophical divides. First, regarding Primitivism, we should note Chalmers's claim that '…to bring consciousness within the scope of a fundamental theory, we need to introduce *new* fundamental properties…' (1996, p.126). EV has some sympathy with this claim, agreeing that the ontology provided by our current conception of the physical cannot account for consciousness. EV just disagrees with Primitivism on what the new properties must be like (Stoljar 2006, p.101). It claims that the new properties are physical rather than experiential properties. This link to Primitivism might give EV some dialectical purchase. Furthermore, it puts the burden of proof on the Primitivist to show why the new properties must be experiential.

Type-A theorists who appeal to the explanatory potential of future science should feel at home with the related, though more sophisticated, position put forward by EV. There is no deep divide to talk across here. Reductionist versions of the Type-A view have a more subtle connection to EV. The Problem of Consciousness involves the physical, the phenomenal and the relation between them. EV and Reductionism agree that looking into the entailment relation is not what will solve the problem. Rather, they claim we must reflect on the relata, and that adequate reflection will undermine the appearance that there is an epistemic gap between them. Where Reductionists assume it is the phenomenal relatum that needs sorting out, EV claims that we must consider our understanding of the physical relatum. As Lockwood explains, '…in reflecting on the relation of consciousness to the *matter* of the brain, philosophers have been apt to take matter for granted, assuming that it is mind rather than matter

that is philosophically problematic' (1989, p.ix). EV challenges that assumption, and encourages Reductionists to take an alternative path to reconciling the two sides of the epistemic gap – a path that does not rest on a dubious analysis of consciousness or risk collapsing into Eliminativism. EV agrees with the Reductionist that we cannot *really* conceive of zombies and inverts, and that Mary would not *really* learn something new. It just disagrees about which side of the apparent epistemic gap is responsible for our confusion. Overall, EV has some potential dialectical force when addressing standard Type-A theorists.

A central assertion of the Type-B view is that claims of conceivability are a limited guide to what is possible. EV agrees with this and holds that when we claim to conceive of zombies or inverts we are guilty of certain standard errors in modal reasoning. Some Type-B theorists claim that conceivability is never a good test for possibility, but EV denies this general claim. Gertler explains that according to EV '[a] local modal fallibilism is all that's required to defuse the problem of experience, on this view; a global modal skepticism is unnecessary.' (2009, p.385) Other Type-B theorists claim that there is something distinctive to consciousness that leads our modal reasoning astray. EV agrees with this, but claims that the source of the error is our ignorance of the physical, not the peculiarities of our phenomenal concepts. Nevertheless, EV has some relevant overlap with the Type-B view.

Of course, in emphasising the various respects in which EV sympathises with existing positions we should not lose sight of its originality. EV offers a quite distinctive account of how we should respond to the problem. Where new variants on existing positions each attempt to plug the holes in a sinking ship, EV offers us a whole new boat. This is important because it is plausible that the right solution will not be achieved by making small adjustments to unpersuasive positions, but rather by pursuing an altogether different kind of strategy.[14]

[14] Stoljar emphasises that one of the respects in which EV differs from standard responses is that it does not assume that 'all the relevant facts are in [and] we just need to think through those facts' (2006, p.143). This is, however, misleading. EV claims we lack the information needed to explain consciousness, but it does *not* claim that we lack the information needed to respond to the Problem of Consciousness. We have, for instance, information about how ignorance leads modal reasoning astray. In this sense, EV is like other positions in claiming that a solution to the problem requires us to think through the facts we

2.2. HISTORICAL PRECEDENT

Stoljar argues at length that there are *historical cases* in which our conceptual ignorance has led to philosophical difficulties analogous to the Problem of Consciousness (2006, pp.123-141). In those cases, an extension of our conceptual repertoire made available a *new type* of fact that allowed the apparent problem to be dissolved. Such cases illustrate that conceptual ignorance can indeed generate an illusion of ontic distinctness. Though this has already been illustrated by the slug thought experiment, finding real-life examples puts EV on firmer footing. They serve to show that *if* we are in a situation analogous to those historical cases, *then* we would wrongly believe that consciousness is onticly distinct from the physical. Of course, this does not give us any reason to believe that we *really are* in such a situation, but the live possibility of our being so encourages us to take EV seriously.

The two historical cases that Stoljar draws on are the explanation of intellectual abilities, as explored by Descartes, and the explanation of chemical bonding, as explored by C. D. Broad. The first of these has the appeal of, like the Problem of Consciousness, being a variant of the mind-body problem. However, Stoljar's depiction of this rests on a questionable interpretation of Descartes' positionFurthermore, it presents a putative case of intermediate-level ignorance. Since we have already concluded that the ignorance hypothesis should be understood as a claim of basic-level ignorance, the example will not be useful to us.

The second example, though more distant from the subject matter of the Problem of Consciousness, offers a more convincing example of the relevant epistemic situation. Broad explains that:

> Oxygen has certain properties and Hydrogen has certain other properties. They combine to form water, and the proportions in which they do this are fixed. Nothing that we know about Oxygen by itself or in its combinations with anything but Hydrogen would give us the least reason to suppose that it would combine with Hydrogen at all. Nothing

---

already know rather than making new discoveries. We just need to be careful to distinguish the facts relevant to the philosophical problem from the facts relevant to the empirical problem.

> that we know about Hydrogen by itself or in its combination with anything but Oxygen would give us the least reason to expect that it would combine with Oxygen at all...Here we have a clear instance of a case where, so far as we can tell, the properties of a whole composed of two constituents could not have been predicted from a knowledge of the properties of those constituents taken separately, or from this combined with a knowledge of other wholes which contain these constituents. (quoted Stoljar 2006, p.135) [15]

Broad thinks that the chemical properties of Oxygen and Hydrogen are not entailed by their mechanical properties. In other words, he thinks that the *chemical* facts are inexplicable in terms of the non-chemical. This leads him to the emergentist conclusion that the chemical facts are distinct from the non-chemical facts. Broad offers reasoning analogous to the knowledge argument here, observing that knowing the mechanical facts is not enough to infer the chemical facts.[16] Broad is plausibly *conceptually ignorant* of the quantum-mechanical properties that we now know explain the chemical properties. Stoljar appeals to McLaughlin's study of emergentism in which he states '…its rise was not due to "philosophical mistakes," nor its fall to the uncovering of such mistakes. ... Advances in science, not philosophical criticism, led to the fall of British emergentism.' (quoted Stoljar 2006, p.140).[17]

The relevance to EV is clear. Broad's conceptual ignorance led to the appearance of an ontological gap between two kinds of property that are in fact seamlessly connected. More specifically, it led him to support something analogous to KA, citing the impossibility of inferring the chemical facts from a complete knowledge of the constituent elements. It is also easy to imagine Broad finding an analogue of CA convincing too. Though Broad's arguments seemed convincing to him and his contemporaries, they were mistaken. Stoljar draws the following lesson from this:

> Just as the chemical argument was plausible to him, so the knowledge argument is plausible to us. Just as it is mistaken to follow the chemical

[15] Interestingly, the motivation for Broad's exploration of the metaphysics of chemistry is to establish analogies with the metaphysics of mind.

[16] A few passages later Broad introduces the notion of a 'mathematical archangel' in an argument recognised as a predecessor to the Knowledge Argument (Stoljar 2006, pp.136-7).

[17] Chomsky claims that Stoljar's account of Broad's '…is somewhat misleading.' (2009, p.199) Chomsky argues that quantum-mechanical discoveries changed the meaning of 'physical facts'. However, he fails to show that the discoveries really did bring about a semantic shift, rather than an epistemic shift. He also seems to overstate how important the concept 'physical' is to Broad's metaphysical conclusions concerning the status of chemistry.

> argument to its conclusion, so it is mistaken to follow the knowledge argument to its conclusion. Finally, just as Broad was ignorant of a type of nonchemical truth relevant to the nature of chemistry, so, too, we are ignorant of a type of nonexperiential truth relevant to the nature of experience. (2006, p.140)

This historical example gives us further reason to believe that EV is worthy of serious consideration.

SECTION 3

WHEN SHOULD WE BELIEVE THE IGNORANCE HYPOTHESIS?

In Section 1 we distinguished EV's two claims: that if the ignorance hypothesis is true the Problem of Consciousness is solved, and that the ignorance hypothesis is in fact true. We have come far enough to see that the first of these claims is plausible. If we are indeed in the kind of situation that the ignorance hypothesis says we are, Physicalism would be true but we would be prone to thinking that there is an epistemic gap between the physical and the phenomenal. Consequently, if you believe that the ignorance hypothesis is true, the case for Primitivism should hold no sway for you. But why should we believe EV's second claim that the ignorance hypothesis is actually true? EV must be able to make a case in favour of the ignorance hypothesis. Justifying the ignorance hypothesis raises a number of issues. In 3.1 I introduce these issues and draw the negative conclusion that certain existing attempts to justify the ignorance hypothesis are inadequate. In 3.2 and 3.3 I offer a constructive account of what is required of EV. Specifically, I argue that we should advocate the ignorance hypothesis only if two conditions can be satisfied. In 3.4 I argue that satisfying these conditions would give us sufficient reason to advocate the ignorance hypothesis.

3.1. A METHODOLOGICAL ISSUE FOR EV

If the ignorance hypothesis is true, it is a contingent truth. It is contingent that the hypothesised physical facts integral to the explanation of consciousness obtain. It is

also contingent that we are conceptually ignorant of those facts. As a contingent truth about our world, the ignorance hypothesis is not the kind of truth we can demonstrate through conceptual analysis alone (Stoljar 2006, p.87). The proponent of EV must provide *evidence* that the ignorance hypothesis actually holds. Without this, EV is merely speculative. We should not demand EV to provide *proof* that the ignorance hypothesis is true (Stoljar 2006, p.141), but we should demand plausible positive arguments for its truth. After all, as Hohwy explains, '…without such an argument, talking about how the problem *could* be resolved if we were in a certain state of ignorance is nothing but idle speculation.' (2005, p.76)

That said, we cannot expect *too much* of EV in this regard. Stoljar explains that it is impossible '…for us to stand outside ourselves and say, "the truths of which we are ignorant are…and…" where the ellipses are filled in by an articulate statement of the truths in question.' (2006, p.87) To do so would clearly involve a 'pragmatic contradiction' (2006, p.87). The target for EV is to give us adequate reason to believe that we are in fact ignorant of a type of physical truth integral to the explanation of consciousness, but without reaching beyond the confines of our proposed epistemic situation to actually provide the truths in question. I will begin by considering Stoljar's attempt to provide such a case, but will conclude that the considerations he raises are too general to constitute a persuasive case in favour of the ignorance hypothesis.

3.1.1. Stoljar's Non-committal Approach

Stoljar recommends an 'abstract' and 'conservative' formulation of EV that is not committed to details about the content of our ignorance (2006, p.121). Accordingly, the considerations he offers in favour of the ignorance hypothesis being true do not involve specific claims about the kind of truth of which we are ignorant, or even about why we are ignorant of them. One thing we can allow from the outset is that it is quite plausible that there are types of physical truth of which we are conceptually ignorant. As Rovane concedes in her critique of McGinn, '[i]t would be foolish to deny that we are subject to significant cognitive limitations' (1994, p.157). Stoljar notes that some basic observations about our position in the world indicate that we should not assume that all physical truths fall within the scope of our conceptual repertoire. After all, our

cognitive abilities have an evolutionary origin and, as Jackson argues, we have no reason to believe that faculties generated in virtue of their survival value are able to penetrate all aspects of the physical universe (Stoljar 2006, p.88). Furthermore, looking at the epistemic limitations of other creatures suggests that we too have epistemic abilities limited by our biology (2006, p.89). Of course, the plausibility of our being ignorant of *some* type of physical truth or other does not give us reason to believe that we are *specifically* ignorant of truths relevant to the explanation of consciousness (Stoljar 2006, p.96). Stoljar makes three main points that he claims constitute indirect evidence for the specific conclusion that the ignorance hypothesis is true. I will consider each point and its limitations in turn.

First, Stoljar argues that the power of the ignorance hypothesis to explain our philosophical predicament regarding consciousness is itself a good reason to believe that the hypothesis is true (2006, p.97).[18] Treating the Problem of Consciousness as a datum, and the various responses to the problem as competing explanations, Stoljar's claim seems to be that an inference to the best explanation lends support to the ignorance hypothesis. In particular, Stoljar notes that the ignorance hypothesis explains our predicament in '…a simple and unified way…', which gives us good reason to believe it is true (2006, p.97).

There are several problems with this argument. The main issue is that it seems to presuppose precisely what it is meant to show – that the ignorance hypothesis is plausibly true. We have already granted that *if* it is true it would explain our philosophical predicament. What we require are arguments that show it *is* in fact true. Its explanatory potential alone cannot achieve this. Imagine a Primitivist theory according to which God bestows consciousness upon us and intervenes in the natural order to make sure our physical behaviour corresponds appropriately to the phenomenal states we have. Despite this theory's capacity to account for our situation in a simple and unified way, it is wildly implausible. As such, the fact that the ignorance hypothesis would explain our predicament does not in itself make that hypothesis plausible.

[18] A similar argument is proposed by Strawson (1994, p.50).

Of course, Stoljar's argument is not just that the ignorance hypothesis is a *possible* explanation of our predicament, but rather that it is the *best* explanation. However, the ignorance hypothesis is only preferable to the competing explanations of our predicament if a case can be made for its plausibility, which is precisely the issue under contention. To present an argument premised on EV being the best explanation thus begs the question against the competing positions, and puts EV in a weak dialectical position when addressing those sympathetic to such positions. Furthermore, an inference to the best explanation must be tempered by a minimal adequacy condition. Even if we could be sure that all the competing positions fail, we can only infer that the ignorance hypothesis is true if it meets that minimal adequacy condition. However, unless a case is made for the plausibility of the ignorance hypothesis it is unclear whether that condition has been satisfied.

Second, Stoljar argues that the ignorance hypothesis '…is a natural conjecture given the fact of our empirical ignorance of conscious experience' (2006, p.105). The claim here is not merely that there are some facts about consciousness that science has yet to account for. It is the much deeper claim that we have '…no framework for thinking about consciousness…' like we do for thinking about other phenomena (2006, p.96). Stoljar claims it is plausible that this deep empirical ignorance involves ignorance of a type of truth relevant to the explanation of consciousness.

The problem with this line of thought is that the fact of our empirical ignorance does not lend adequate support to the ignorance hypothesis. It lends plausibility to the hypothesis in the following modest sense: if we had rich empirical knowledge of a certain domain, it would be implausible to claim that we are conceptually ignorant of a type of truth in that domain. The inverse, however, does not hold: a lack of empirical knowledge need not indicate conceptual ignorance. Perhaps the unknown empirical truths about consciousness are all truths available within our current conception of the physical. Why should we believe that they are a different type of truth to familiar biological and cognitive truths? Of course, one could claim that the familiar types of truth are inevitably irrelevant to the explanation of consciousness, so the unknown truths *must* be beyond our current conception. This, however, would collapse into Stoljar's first point that we should believe the ignorance hypothesis because it solves

the Problem of Consciousness. This would merely be a restatement of EV rather than an independent point in favour of the ignorance hypothesis.

Stoljar's third consideration in favour of the ignorance hypothesis is that it has considerable historical precedent. We have already discussed such precedent, but does this really lend support to the ignorance hypothesis? The precedent shows that *if* you are in the kind of epistemic situation proposed by the ignorance hypothesis *then* you would find yourself in the kind of philosophical predicament we find ourselves in with consciousness. But why should this encourage us to believe that we are in fact in that epistemic situation with regards to consciousness? Surely it would be foolhardy to turn the conditional the other way and claim that *if* you find yourself in that kind of predicament *then* you are probably suffering from the proposed ignorance? We still need positive reason to believe the ignorance hypothesis is true.

Overall, Stoljar's points fail to provide what is needed. In Stoljar's defence, he may have had a more modest goal in mind when presenting these points. They plausibly succeed in achieving his stated aim of 'rendering it [the ignorance hypothesis] believable' (2006, p.87). Nevertheless, we should only advocate EV if we have adequate reason to believe that the ignorance hypothesis is actually *true* rather than just a live possibility. A critic of EV might concede that the hypothesis is a genuine option, but regard it is an unsubstantiated piece of speculation that we have inadequate justification for believing. We can capture this with a distinction made by Bennett: 'p is weakly plausible if we are not certain that ~p [and] p is strongly plausible if we have positive reason to believe p.' (2009, p.772) Stoljar's arguments may undermine the critic who claims to be certain that the ignorance hypothesis is false. They will not, however, persuade the critic who requests positive reason to believe that the ignorance hypothesis is true. Unless arguments can be provided that address such a critic, the case for EV is incomplete.

3.1.2. Overreaching

The lesson we learn from evaluating the points raised by Stoljar is that general considerations about the possibility of our being ignorant are inadequate. EV needs

positive arguments in favour of the *specific* claim that there are physical properties beyond our current conception that are integral to the explanation of consciousness. However, there is a threat inevitably associated with attempts to provide such arguments. In trying to say something about the content of our ignorance, it is easy to make speculations that go beyond what we are really in an epistemic position to justify. If an argument *overreaches* in this way then a critic will still be able to object that EV tells a coherent story about our situation regarding consciousness, but does not give us adequate reason to believe that the story is true. To illustrate this issue I will briefly consider a position that is guilty of such overreaching: McGinn's 'mysterianism'.[19] First though, I will provide a framework for understanding how details of the ignorance hypothesis could be filled in.

Given that an advocate of the ignorance hypothesis cannot, without self-contradiction, state the truths that it claims are unknown, how can it provide any details about the content of our ignorance? There are three dimensions along which information about the unknown physical properties might be provided. First, claims could be made about the relationship of the unknown property to the physical properties of which we do have a conception. Even without acquiring the hypothetical missing concept, we might be able to draw conclusions about the place of the unknown properties in the familiar physical world, and about ways in which it is like or unlike known physical properties. Second, claims could be made about the relationship of the unknown properties to our cognitive faculties. After all, we are owed an account of *why* these physical properties are currently beyond our grasp. Third, claims could be made about the relationship of the unknown properties to consciousness. EV is committed to them playing an integral role in the explanation of consciousness, but there is space to say more about that role. For instance, is the instantiation of the unknown property necessary or sufficient for the instantiation of phenomenal properties?

Of course, all three of these dimensions of detail are intimately related; conclusions drawn regarding one dimension will have implications for the others. With this framework in mind, we can offer a quick outline of McGinn's distinctive

[19] The term 'mysterianism' was coined by Flanagan (1992) as a label for McGinn's position.

formulation of the ignorance hypothesis. McGinn argues that there is a 'property *P*' that is responsible for consciousness and which is permanently beyond our grasp (1989/2004).[20] On the first dimension he suggests, for instance, that the states responsible for consciousness do not '…have the marks of full-blown objectivity…' (2004, p.91) and that they go '…beyond the spatial properties recognised in physical science' (2004, p.104). On the second dimension, he argues that our concept-forming procedures are incapable of producing a concept of *P*. The key step in his argument is that our theoretical concepts are based on perceptual data, so can only reveal properties akin to perceptible properties. Since consciousness is *im*perceptible, no property akin to perceptible properties can account for its existence. On the third dimension, McGinn seems to hold that property *P* is entirely responsible for consciousness rather than being an unknown piece in an otherwise accessible puzzle (e.g. 2004, p.59). He also suggests that the mechanism by which the unknown properties generate consciousness is 'non-combinatorial', but argues for the Chomskyan conclusion that we are only capable of understanding 'combinatorial' explanations in which discrete atoms combine in a law-like manner (1994 and 2004, Ch.8).

Without offering any detailed evaluation of McGinn, I will note some of the serious objections that have been raised against his fleshing-out of the ignorance hypothesis. Whiteley argues that property *P* being heterogeneous to familiar physical properties raises worries about it having physical causes or effects, which in turn raises worries about the causal status of consciousness (1990, p.394). The cognitive psychology McGinn uses to make his case for the epistemic inaccessibility of *P* is questionable. Rovane (1994) and Kukla (1995) find counter-examples to McGinn's model of our concept-forming procedures. Kirk (1991) and Rovane (1994) argue that McGinn fails to recognise the possibility of our different cognitive capacities working together to overcome the limitations that they have in isolation. Stoljar (2006, pp.92-93) challenges McGinn's notion that theoretical concepts are restrained by perceptual data. On a strong reading of McGinn's claim, it is simply false, and on a weaker

[20] McGinn does not claim that property *P* is a physical property, so does not technically advocate the ignorance hypothesis as a defence of Physicalism. His position is nevertheless close to EV, so is worthy of some attention.

reading, it fails to generate the conclusion that no theoretical concept could be relevant to the explanation of consciousness. Regarding McGinn's claims about the explanatory role of property *P*, Sacks expresses doubt that it is possible for any property to perform the proposed explanatory task, leaving McGinn no better off than a standard Physicalist (1994, p.34). Furthermore, Rovane argues against the view that we cannot understand 'non-combinatorial' modes of explanation (1994).

The lesson we learn from this brief examination of McGinn's position is that filling in the ignorance hypothesis is no mean feat. A theme in the objections to McGinn is that he is *overreaching* – that he is telling a story that *might* be true but for which we are far from having sufficient evidence. If we are going to make an adequate case for the ignorance hypothesis, we will have to avoid such overreaching. Another lesson we can take away is that detailing our epistemic situation is not always *relevant* to the plausibility of the ignorance hypothesis. Stoljar notes that McGinn's claim that our ignorance is *chronic* is irrelevant to the philosophical problem (2006, p.93). The question is whether there are unknown truths that are integral to the explanation of consciousness, and whether or not our ignorance of them is *permanent* has no bearing on that claim. Against Stoljar, I have argued that we need focused arguments in favour of the ignorance hypothesis but, with Stoljar, I can conclude that we should not be distracted by details that are irrelevant to the plausibility of the hypothesis. In attempting to flesh out the ignorance hypothesis, we must be careful to avoid bells and whistles that distract from the real question and which increase the threat of overreaching.

We know that the plausibility of EV depends on a case being made for the ignorance hypothesis. So far we have reached some conclusions about what such a case must avoid. It should not rely on general considerations about the possibility of our being ignorant as they fail to lend support to the specific conclusion that we are in fact in the epistemic situation described by the ignorance hypothesis. Nor should it overreach: it should not attempt to state the truths of which we are ignorant, nor should it speculate about the content of our ignorance without sufficient evidence, nor should it embellish the ignorance hypothesis with details irrelevant to its plausibility. The task now is to reach some more positive conclusions about what the case for the

ignorance hypothesis must achieve. I will introduce two conditions that a case in favour of the ignorance hypothesis must satisfy.

3.2. THE RELEVANCE CONDITION

3.2.1. The Condition

If EV is going to persuade us to believe that the ignorance hypothesis is true, it will have to address doubts about the very possibility of some unconceived physical property performing the proposed role in the explanation of consciousness. Advocates of the epistemic gap claim that *all* physical truths are, in principle, unable to entail the phenomenal truths. The burden of proof is then on EV to give us reason to believe otherwise. Trogden captures this challenge to EV with reference to the story of the slugs:

> Though the slugs aren't in a position to perceptually detect pie piece shapes, as sophisticated cognizers they could see how truths about nondetected nontriangular tiles are the sorts of considerations that could render the circular truths intelligible in a world that is fundamentally noncircular. (2009, p.271)

The objection to EV is that, in this respect, our situation regarding consciousness is quite *unlike* the slug's situation regarding circles. We cannot see how any physical truth could be relevant to consciousness. Trogden dubs this the 'really-not-having-a-clue feature' of our situation, and it is this feature that justifies resistance to the ignorance hypothesis (2009, p.272). The reason we really don't have a clue about how a physical truth could perform the requisite explanatory role is that there are a priori obstacles that cast doubt on the possibility of them doing so (see Hohwy 2005, p.77). Before we move on to discuss those a priori obstacles, we can state the challenge faced by EV:

> ***The Relevance Condition:*** The ignorance hypothesis should be advocated only if we have adequate reason to believe that unconceived physical properties could evade the a priori obstacles to a physical explanation of consciousness.

Clearly the satisfaction of this condition will involve giving some details about the content of our ignorance – details that lend support to the possibility of the unknown physical properties being relevant to the explanation of consciousness. Does this condition unreasonably demand EV to state the truths of which we are ignorant in order to demonstrate their relevance? No – in principle, one could show that an a priori consideration does not apply to a certain kind of property without having a concept of that property. To do this, one would have to know *something* about the property in question, but this is consistent with conceptual ignorance. What, then, are the a priori challenges to the relevance of any unknown physical property? This is where the –tivity and –trinsicality gaps return to the stage.

3.2.2. The Ignorance Hypothesis and the –tivity Gap

The epistemic gap is not just based on a brute intuition that no physical state could ever entail a phenomenal state. It is based on two deeper conceptual gaps concerning what *kind* of state physical states are, and what *kind* of state phenomenal states are. Just as these conceptual gaps can be used as vetoes against the hypothesis that progress in science will close the epistemic gap, so too can they be used as vetoes against the hypothesis that acquiring new physical concepts could close the epistemic gap.

The challenge to the ignorance hypothesis presented by the –tivity gap is captured by Papineau:

> Stoljar is here placing strong demands on the content of our ignorance. It must be such that, if only we knew the relevant non-experiential facts, this would render zombies inconceivable. However, it is not clear that any non-experiential facts could play this role. By their nature, non-experiential facts would seem to be third-personal, objective, and non-perspectival, while experiential facts are first-personal, subjective, and perspectival. It is hard to see how knowledge of the former could automatically render the absence of the latter inconceivable. (2007)

The same concerns apply, *mutatis mutandis*, to EV's account of KA. In response, the proponent of EV could claim that the unknown physical properties are actually subjectivity-involving. This route, however, would concede too much to Primitivism. EV

promises to protect the view that experiential states are entailed by non-experiential states, so to posit unknown experience-involving properties would be to renege on that promise. Alternatively, EV could argue that there is a *third* category of property that is neither fully objective nor fully subjective. If the unknown properties fall into this third category, they will not face the problem of accounting for subjectivity in objective terms, but nor will they concede to the Primitivist that subjectivity is onticly basic. We have already seen that McGinn sympathises with this kind of neutral claim (2004, p.91). However, it is doubtful that the notion of such a third category is coherent, and doubtful that its introduction would really help account for the subjectivity of phenomenal states.

The only option for EV is to maintain that there are robustly objective truths beyond our current conception that are suited to bringing about subjective states. Indeed, this is precisely the route that Stoljar takes. To cast doubt on the –tivity gap, Stoljar invites us to consider the proposition '[i]f John is a number, then he is not in pain' (2006, p.157). Stoljar's claim is that the antecedent is an objective fact that entails a subjective consequent. This is intended to act as a counter-example to the putative principle that objective facts cannot entail subjective facts. If that general claim can be undermined, the objectivity of an unknown property need not count against its explanatory relevance to subjective phenomenal states.

One objection to this argument is that the epistemic gap is concerned with the inexplicability of positive phenomenal facts – facts that concern the *presence* of subjective states. The example provided shows, at best, that something objective can entail the *absence* of something subjective. It is not clear why this should assuage doubts about the possibility of physical states entailing the *presence* of subjective states (see Papineau, 2007). Furthermore, the entailment in the example holds because being a number conceptually *excludes* being in pain. Again, it is unclear how entailment by exclusion has any bearing on the kind of positive entailments required for a physical explanation of subjective states (see Hardcastle, 2008 and Robinson, unpublished). Overall, the –tivity gap presents a serious challenge to the ignorance

hypothesis.[21] The Relevance Condition can only be satisfied if this challenge is addressed.

3.2.3. The Ignorance Hypothesis and the –trinsicality Gap

Unsurprisingly, a parallel challenge to the ignorance hypothesis is provided by the –trinsicality gap. This challenge is captured by Alter. He first identifies the structure and dynamics thesis (SDT): 'There are experiential truths that cannot be deduced from truths solely about structure and dynamics.' (2009, p.760) Our reason for accepting SDT is that phenomenal qualities are non-structural properties. Alter then goes on to present the following simple argument against the ignorance hypothesis:

> 1. The ignorance hypothesis undermines CA & KA only if it undermines SDT.
>
> 2. The ignorance hypothesis does not undermine SDT.
>
> 3. Therefore…the ignorance hypothesis does not undermine CA & KA. (2009, p.760)

Again, Stoljar attempts to undermine the claim that there can be no entailment from extrinsic/structural properties to intrinsic/non-structural properties. He presents the example of Mr. & Mrs. Spratt. 'Being married' is a relational property of Mr. Spratt and a relational property of Mrs. Spratt. But, Stoljar claims, '…being married is an intrinsic property of the pair…' (2006, p.152). Since this intrinsic property can be derived from the relational properties of each of the Spratts, it is not the case that there can be no entailment from extrinsic properties to intrinsic properties.

The key problem with Stoljar's argument is that even if we allowed that 'being married' is an intrinsic property of the Spratts, it is not the *kind* of intrinsic property that drives the –trinsicality gap. Pereboom's recent insights allow us to identify what's wrong with Stoljar's example. Pereboom stipulates that a property is *comparatively* intrinsic if it is an intrinsic property of X that '…reduces to parts of X having purely extrinsic properties.' (2011, p.94) 'Being married' comes out as a comparatively intrinsic property of the Spratts. The intrinsic qualities that characterise phenomenal

[21] A complication here is that Stoljar does not follow our characterisation of subjectivity as the existence condition of a phenomenal state (Chapter 1). Nevertheless, the same points still hold.

states, on the other hand, do not appear to be comparatively intrinsic. The reason that qualitative redness appears to be physically inexplicable is precisely that it resists analysis into constituent parts, and that its essence resists any purely structural characterisation. In Pereboom's terms, phenomenal qualities are *absolutely* intrinsic properties. The possibility of an entailment from extrinsic properties to comparatively intrinsic properties does nothing to assuage doubts about the possibility of an entailment from extrinsic properties to absolutely intrinsic properties. We will have more to say on intrinsicality and related concepts in the next chapter, but it is already clear that Stoljar faces a significant challenge here.[22]

Overall, Stoljar's brief attempt to undermine worries about subjectivity and intrinsicality does not achieve what it should. As Hardcastle objects, '…Stoljar does not appear to take those who object to his perspective seriously enough' (2008, p.275). The –tivity and –trinsicality gaps play an integral role in the Problem of Consciousness, and Stoljar does not show adequate concern for them.

3.3. THE INTEGRATION CONDITION

3.3.1. The Condition

The Relevance Condition concerns the relationship between consciousness and the hypothetical unknown physical truths. The plausibility of the ignorance hypothesis also depends on the relationship between the hypothetical unknown physical truths and the *known* physical truths. We already have a rich picture of the physical world, and the ignorance hypothesis must dovetail with our existing knowledge of it. Our knowledge of the physical world is a posteriori, so this condition contrasts with the a priori challenge presented by the Relevance Condition.

> ***The Integration Condition****:* The ignorance hypothesis should be advocated only if we have adequate reason to believe that the holes in

[22] Stoljar also attempts to resist the –trinsicality gap by casting doubt on the intrinsicality of phenomenal qualities (2006, pp.149-151). As Coleman comments (2007, p.830), Stoljar's appeal to the diaphanousness of experience misses the point. I will explain why this Representationalist move fails in Chapter 5.

> our current conception of the physical offer a suitable place for properties relevant to the explanation of consciousness.

We have already accepted the likelihood that there are types of physical truth of which we are ignorant. It is one thing to claim that we suffer from conceptual blind-spots, and quite another to claim that any such blind-spot is an appropriate home for properties relevant to the explanation of consciousness. The conclusion that unconceived properties *could* close the epistemic is worthless unless those properties can be suitably accommodated in our world-view. EV must show that there is an appropriate *home* for the proposed unknown properties within the physical world. What criteria determine whether or not an established blind-spot is plausibly occupied by properties relevant to the explanation of consciousness?

First, EV's answer to the *Relevance Condition* will doubtless bring along commitments about the nature of the unknown physical properties. Once it has been argued that unknown physical properties could plausibly evade the –tivity gap and the –trinsicality gap, there remains the task of integrating these hypothetical properties into our understanding of the physical world. A conceptual blind-spot is only relevant if we have positive reason to believe that the kind of property that occupies that blind-spot is the kind of property that could close the epistemic gap.[23]

Second, the blind-spot must integrate with what we already know about how consciousness fits in to the physical world. For instance, we have reason to believe that the properties responsible for consciousness are instantiated in the brain. As such, an instance of conceptual ignorance is only relevant if the unknown properties are instantiated by the brain. Of course, there is some flexibility here. Given our failure to explain consciousness in physical terms, we cannot take our presuppositions about the explanatory basis of consciousness for granted. That said, the onus is on EV to show that such presuppositions are inaccurate. For example, if our ignorance is *not* of properties instantiated by the brain, EV had better be able to tell a plausible story about why the brain *appears* to be the seat of consciousness.

[23] This criterion will be easier to appreciate in the next chapter when we attempt to specify the nature of the unknown properties.

Third, a related concern is that the ignorance hypothesis must integrate with what we already know about the *causal status* of consciousness. We know that conscious states have physical effects. More specifically, we know that our conscious states influence our bodily states. The causal status of conscious states is fixed by the causal status of the properties responsible for that state. As such, a conceptual blind-spot is only suitable if it is plausibly occupied by properties that play an appropriate causal role. Again, there is some room for EV to challenge our presuppositions about the causal characteristics of conscious states, but not a lot. After all, it is causal considerations like these that led us to reject Primitivism.[24]

3.3.2. Ignorance and Knowledge

The satisfaction of each of these criteria must work in tandem with a more fundamental requirement. An advocate of EV can only integrate hypothetical properties into our understanding of the physical world if they have first triangulated a clear conceptual blind-spot. We have to be sure that we are suffering from a specific case of conceptual ignorance before we can be sure that the properties of which we are ignorant are suitably placed to perform the proposed explanatory role. In attempting to demonstrate that we are suffering from a conceptual blind-spot, there are two key considerations that must be respected.

First, a claim of ignorance must be able to account for how we have managed without the proposed missing concept. Prinz makes a representative claim about our knowledge of the brain:

> …I think we can understand the brain quite fully. The mind–body problem does not stem from an inability to see the brain in all its glory. There is no hidden property that escapes detection when we turn on our magnets, implant electrodes or stain chunks of brain. (2003, p.116)

The ignorance hypothesis is committed to Prinz being wrong, but it must be able to accommodate the *plausibility* of Prinz's claim. A parallel point can be made in the context of physics. Strawson claims that physics does not '…seem to have any obvious

[24] Dauer (2001) and Brueckner & Beroukhim (2003) argue that McGinn's position has epiphenomenalist implications that undermine its plausibility.

promising gaps or valences where conceptual extension could possibly bring in radically new predicates …' (1994, p.95). Why are our existing physical theories so good at explaining phenomena other than consciousness if they are working with an impoverished conceptual repertoire? Why don't we *know* the proposed unknown properties? How have they managed to stay hidden?

Second, a claim of ignorance must be able to show that the proposed unknown properties are *actual*. Consider the possibility that we have no concept of certain divine properties. This conceptual limitation only generates ignorance if there is an *actual divine being* that instantiates the properties of which we have no conception. Similarly, our lacking some physical concept only generates ignorance if such properties are actual. EV must be able to show that our blind-spot is occupied – it must provide evidence of the instantiation of these unconceived properties.

These last two points are significant challenges, but the real trouble for EV comes when we consider their relationship to one another. If the unknown properties make a manifest difference to physical events of which we have knowledge, we have evidence of the presence of those properties, and so reason to believe that those properties are actual. However, in this situation it is hard to account for why we haven't noticed our ignorance, or for how we have failed to form a conception of the properties in question. Alternatively, it could be held that unknown properties make no manifest difference to known physical events. This makes it easy to explain why we haven't noticed that we suffer from a limited conceptual repertoire. However, in this situation it is hard to justify the conclusion that such properties are actual. What evidence could we have for their existence if they make no manifest difference to known physical events? Finding a way to respect both criteria at once thus poses a significant challenge to advocates of the ignorance hypothesis.

To summarise, the Integration Condition demands a positive account of the place of the hypothetical unknown properties in the physical world. This requires the identification of a blind-spot in our current conception of the physical, with all the problems that entails. It then requires positive reason to believe that the blind-spot is occupied by properties that a) respect any conclusions reached in response to the Relevance Condition, b) respect what we already know about the physical basis of

consciousness and c) respect what we already know about the causal status of consciousness.

3.4. ARE THERE ANY FURTHER CONDITIONS?

I have argued that the ignorance hypothesis should be advocated only if the two conditions above can be satisfied. I now want to argue that if those two conditions can be satisfied, we should adopt the ignorance hypothesis. In other words, if those two challenges can be met then EV should be advocated. To justify this claim, I will consider some objections that have been raised against the ignorance hypothesis. The main objections have already been captured by the two conditions, but there are some further considerations that must be addressed. I argue that these objections are either misguided or can be absorbed into the two established conditions. This justifies the conclusion that we should not place any further conditions on EV.

3.4.1. The Coherence of Conceptual Ignorance

Kriegel argues that '…it is incoherent to suppose that a conceptual scheme is powerful enough to frame a problem without being powerful enough to frame its solution' (2003, p.179). Given that we can frame the Problem of Consciousness, we must, therefore, have the concepts with which to resolve it. Of course, one could deny that we *can* frame the problem, but this would be to shun EV in favour of Chomskyan skepticism. Kriegel supports his conclusion with the following example: 'One cannot be said to understand the question "What is John's weight?" if one does not understand the meaning of "John weighs 150 pounds." Understanding a question is thus necessarily coupled with understanding its possible answers.' (2003, p.184) He goes on to point out that '…the only logic of questions we have today is founded on the individuation of questions in terms of their possible answers...' (2003, p.185) He then asks '…how could we formulate the problem--the *right* problem--without deploying the missing concept?' (2003, p.186).

This objection to EV is misguided. First, Kriegel moves too easily from 'philosophical problems' to 'questions'. Perhaps philosophical problems are not best

understood as questions, and if they are questions, they could easily be special cases quite disanalogous to questions about John's weight. Assuming that philosophical problems are questions, Kriegel still faces difficulties. We can agree that understanding a question involves understanding what it would take for something to be an *answer* to that question. However, Kriegel is making the much bolder claim that we must be able to frame *each of the possible answers*. This is clearly false. Imagine someone who can only perceive on a slightly limited spectrum, and consequently has no concept of red. Such a subject would be able to understand the meaning of the question 'what colour are post boxes?', despite her inability fully to grasp the answer to this question. This is because she has the concept 'colour' (without having concepts for *all* the colours) and the meaning of the question simply concerns colour-as-such. Kriegel's claim that we must be able to entertain all possible answers to a question is motivated by an appeal to our 'only' logic of questions. This move is unpersuasive, since being the *only* logic does not guarantee being a *good* logic.

Dennett (1991b) raises a similar objection to Kriegel when he argues that our capacity to formulate the problem is itself a sign that we are not conceptually ignorant. He notes that a monkey has no concept of an electron, but observes that the monkey is in no way perplexed as a result of its ignorance. Since we *are* perplexed, we are not in an epistemic situation analogous to the monkey's.

Dennett is right that an appeal to ignorance must be able to account for our philosophical perplexity. To say that our ignorance is just like that of the monkey would indeed be unhelpful. However, hopefully we have established that conceptual ignorance *can* explain our philosophical predicament. To deny this would require arguments to be provided against much of the work we have done so far, including the story of the slugs and the case of historical precedent. The claim was never that conceptual ignorance is *sufficient* to generate philosophical puzzles analogous to the Problem of Consciousness, so EV is not committed to the monkey's conceptual ignorance being accompanied by philosophical perplexity. Overall, neither Kriegel nor Dennett's objections have any force.

3.4.2. The Overgeneration Problem

Karen Bennett (2009), in a commentary on Stoljar, argues that if EV is an appropriate response to the Problem of Consciousness, then parallel responses to a whole range of philosophical problems should also be appropriate. Since it is implausible that all such problems can be solved by hypothesising that we suffer from conceptual ignorance, we should not accept EV in the context of consciousness. She explains:

> Here is the general recipe, for a target argument that has a series of premises P1 through Pn:
>
> a) Argue that it is plausible to suppose that we are ignorant of the truth (and perhaps the content) of some proposition p, and
>
> b) argue that our ignorance of p undermines our reason for believing some Pi. (2009, p.772)

Bennett then asks whether this argument requires *p* to be weakly or strongly plausible. If EV only requires our ignorance to be weakly plausible, then we have an absurdly easy response to any number of philosophical problems. However, if our ignorance has to be *strongly* plausible, the problem is that Stoljar's arguments fail to provide strong plausibility to the ignorance hypothesis.

We can simply regard this objection as another reason to set our standards higher than Stoljar. Our two established conditions demand *strong* plausibility. Bennett is probably right that requiring only weak plausibility would mean EV overgenerates philosophical solutions. This just makes it all the more imperative to insist on strong plausibility. As such, Bennett's concerns are already captured in our two conditions.

3.4.3. Relocating the Mystery

Some critics have worried that even if consciousness *is* the result of unconceived properties, this metaphysical entailment will still be mysterious. Sacks argues:

> …given a blind spot, we can *speculate* that there is no more than a problem of access, but since there is that problem of access, it becomes more difficult to establish that what is going on at that spot is not itself problematic. And the more extensive such cognitive closure, the more this difficulty will impinge. (1994, p.32)

In other words, in making a claim of conceptual ignorance EV cannot justify the conclusion that the unknown properties close the epistemic gap. If we suffer from conceptual ignorance, we cannot say one way or the other whether phenomenal properties are primitive.

Gertler makes the related point that '…to say that our own situation is a symptom of ignorance is not, of course, to say that we are ignorant of some fact incompatible with primitivism, as the ignorance hypothesis requires.' (2009, p.383) Perhaps the unknown properties are phenomenal properties. Since EV cannot rule out this possibility, it cannot make a convincing case against Primitivism.

Both of these objections appear to take an inappropriate view of where the burden of proof lies. The plausibility of EV does not depend on it explicitly demonstrating that Primitivism is false. We have independent reason to doubt Primitivism, and basic principles of parsimony mean we should only adopt Primitivism if all serious non-Primitivist options are ruled out. The objections could be read as making the stronger claim that no non-Primitivist version of the ignorance hypothesis is true. But this concern is already captured by the Relevance Condition, as it demands positive reason to believe that the case for Primitivism cannot be wielded against EV.

CONCLUSION

We now understand what EV is, how it attempts to undermine the case for Primitivism, and why it is a promising and attractive proposal. We have concluded that if the ignorance hypothesis is true, then the Problem of Consciousness is solved. Now we have concluded that if the Relevance Condition and Integration Condition can be satisfied, then we should believe that the ignorance hypothesis is true. Therefore, if an account of our ignorance is provided that satisfies those two conditions, then the Problem of Consciousness is solved.

CHAPTER 4

THE RUSSELLIAN IGNORANCE HYPOTHESIS

This chapter explores a prominent way of filling in the ignorance hypothesis, notably associated with Russell (1921/1927). The Russellian Ignorance Hypothesis (RIH) consists of two claims: i) that we are conceptually ignorant of the intrinsic properties of physical entities, and ii) that these properties are integral to the physical explanation of consciousness. In the first three sections of the chapter I argue that a version of 'i' is true. In the fourth and final section I evaluate 'ii', applying the Relevance and Integration Conditions on EV established in the previous chapter. I argue that RIH goes a long way towards satisfying both conditions, but ultimately falls short on the Relevance Condition due to its failure to address the –tivity gap. Consequently, RIH does not offer a viable solution to the Problem of Consciousness. I also argue that since RIH is the most promising version of the ignorance hypothesis, it is likely that *no* version of the ignorance hypothesis can meet our two criteria. In the following chapters I will show how RIH, and with it the Epistemic View, can still be deployed as an essential part of a more complex response to the Problem of Consciousness.

SECTION 1

INTRODUCING INSCRUTABILITY

A number of philosophers, including Russell, have argued that we have no conception of the intrinsic nature of physical entities. Foster (2008) calls this the 'inscrutability' of matter. We can refer to the proposed unknown intrinsic properties as 'inscrutables'. The case for inscrutability is distinct from concerns surrounding the metaphysical

status of the phenomenal, so for the time being we can put consciousness aside.[1] My argument for inscrutables – the Inscrutability Argument (IA) - runs as follows:

> IA1) We have epistemic access to physical objects only via how they affect us.
>
> IA2) Physical objects have absolutely intrinsic properties.
>
> IA3) If we have epistemic access to physical objects only via how they affect us, then we cannot form a transparent conception of the absolutely intrinsic properties of objects.
>
> IA4) Therefore we cannot form a transparent conception of the absolutely intrinsic properties of physical objects.[2]

I will elucidate and defend this argument in several stages. In this section, I will clarify the intrinsic/extrinsic dichotomy and expand on the notion of *absolute* intrinsicality that I introduced in Chapter 3. In Section 2 I will argue for IA's three premises. In Section 3, I supplement the case for IA2 by defending it against the 'Pure Structuralist' claim that physical entities do not have absolutely intrinsic properties. Ultimately I conclude that we are indeed conceptually ignorant of the intrinsic nature of physical entities.

1.1. THE INTRINSIC/EXTRINSIC DISTINCTION

We have already discussed the distinction between intrinsic and extrinsic properties in the context of the –trinsicality gap (Chapter 1, Section 3.2.2). This distinction must be developed further in order to appreciate the argument for inscrutables. The distinction between intrinsic and extrinsic properties is a distinction between those properties a thing has in virtue of itself, and those that it has in virtue of how things stand in the world beyond itself. As Harris notes, '[t]he intrinsic/extrinsic distinction is generally taken to be mutually exclusive and jointly exhaustive of the domain to which it applies.' (2010, p.467) That is, it is thought that every property is either intrinsic or

[1] When Montero introduces the term 'inscrutables' (2010, p.74) she defines them as being responsible for consciousness. My use of the term leaves it open whether or not they have this explanatory role.
[2] The contrast between transparent and opaque ways of knowing is outlined in Chapter 3 (Section 1.1.2) and will be re-examined later in this chapter.

extrinsic and no property is both. The domain in question is that of properties: extrinsicness and intrinsicness are properties of properties.[3]

Socrates's height is plausibly an intrinsic property of Socrates.[4] His height has ramifications for the relations in which he might stand and presumably has an explanation that involves causal relations to his parents' genes. Nevertheless, it is a property that *consists in* something internal to Socrates. By contrast, Socrates's property of 'being shorter than Simmias' is an extrinsic property, depending as it does on how things are in the world beyond Socrates. This is still a property of Socrates himself, but it consists in a relation to a distinct individual. This example offers a useful starting point for an exposition of the intrinsic/extrinsic distinction. Can we go further and offer a *reductive analysis* of the distinction? Attempts to provide such an analysis have consistently failed, but this should not make us suspicious of the distinction. As Seager argues, '…the concept of the intrinsic properties of an object seems intuitively intelligible, despite the difficulties philosophers have in spelling it out precisely in non-question begging terms' (2006, p.130). Putting the project of *defining* the distinction to one side, we can still establish some useful tools for determining whether a property is intrinsic or extrinsic.

Relationality offers an illuminating way of fleshing out the intrinsic/extrinsic distinction.[5] Roughly, properties that depend on a relation (or relations) to a distinct individual (or individuals) are extrinsic, while intrinsic properties are independent of such relations. To test whether a property is relational, we should consider whether an individual would still instantiate that property if it were the only individual in the world. In such a 'lonely' world, there are no distinct individuals with which to stand in relations, and so no potential to possess relational properties. As such, if a property instantiation is compatible with the loneliness of its bearer, that indicates it is an intrinsic property. For instance, Socrates would plausibly still have his height in a lonely world, but he would not have the property of being shorter than Simmias. However, compatibility/incompatibility with loneliness is not quite equivalent to the

[3] A complication with this second-order property account is explored by Humberstone (1996). There is a 'local' sense of intrinsic/extrinsic that describes *ways of having* a property rather than attributing a second-order property. This alternative use will not concern us.
[4] This is an adaptation of an example from Van Cleve (2002).
[5] Francescotti (1999) offers an interesting account in terms of what he calls 'D-relationality'.

intrinsic/extrinsic distinction. For instance, 'not being five feet from a rhododendron' (Humberstone, 1996) and 'being a lonely object' (Lewis 1983, p.199) are compatible with loneliness but are extrinsic properties. Nevertheless, loneliness-compatibility offers a handy diagnostic tool that points us in the right direction.

Another useful test of intrinsicality is 'duplicate-invariance'. Your extrinsic properties depend on how the world beyond you stands i.e. on which possible world you are in. Your intrinsic properties do not, so would remain constant regardless of which modal neighbourhood you occupy. In light of this Lewis proposes that '[i]f something has an intrinsic property, then so does any perfect duplicate of that thing; whereas duplicates situated in different surroundings will differ in their extrinsic properties' (1983, p.197). This does not quite work as a *definition* of the distinction: it faces some tricky counter-examples, and the notion of a 'duplicate' is hard to characterise without circular reference back to intrinsicality (see Seager 2006, pp.129-30 and Francescotti 1999, p.593). Nevertheless, duplicate-invariance offers another handy diagnostic tool.

If a property passes both the loneliness-compatibility and duplicate-invariance tests, then we should regard it as intrinsic. If it does not, we should regard it is intrinsic. Strictly speaking this two-part test is not infallible, but it is quite sufficient for our purposes.

The extrinsic and intrinsic properties of an individual are intimately connected. Drawing on Pereboom (2011), we can offer some more refined distinctions that enable us to map these connections. Some extrinsic properties place constraints on the intrinsic properties of their bearer. For instance, Socrates's extrinsic property of being shorter than Simmias is constituted in part by his intrinsic property of being the height he is. Other extrinsic properties place no such restraints. For instance, 'being one among many' tells us nothing about the intrinsic properties of its bearer (Pereboom 2011, p.93). To capture this, Pereboom introduces a notion of *pure* extrinsicality:

> P is a *purely extrinsic property* of X just in case P is an extrinsic property of X and P has no intrinsic aspects. (2011, p.93)

Just as some extrinsic properties place demands on the intrinsic properties of their bearer, some intrinsic properties place demands on their bearer's extrinsic properties.

As previously discussed, 'being married' is perhaps an intrinsic property of the pair Mr. and Mrs. Spratt. However, the pair instantiates this intrinsic property only in virtue of its constituent parts each having the extrinsic property of being married to their spouse. This kind of situation leads Pereboom to draw the following distinction:

> P is an *absolutely intrinsic* property of X just in case P is an intrinsic property of X, and X's having P does not reduce to parts of X having purely extrinsic properties…
>
> P is a *comparatively* (or a *relatively*) *intrinsic* property of X just in case P is an intrinsic property of X, and X's having P reduces to parts of X having purely extrinsic properties. (2011, pp.93-94)[6]

These notions can also be defined in terms of whether an object's possession of an intrinsic property is derivable a priori from a proposition describing the purely extrinsic properties of its parts (Pereboom 2011, p.94). IA – the Inscrutability Argument – specifically claims that we are ignorant of the *absolutely* intrinsic properties of physical objects.

1.2. CLARIFYING THE DISTINCTION

So far we have made some positive claims about what the intrinsic/extrinsic distinction is. Now I will make some negative claims about what the distinction is *not*. The intrinsic/extrinsic distinction is often conflated with other conceptual dichotomies, so it is worth explicitly ruling those conflations out. I will briefly outline three mistakes that are liable to be made regarding the distinction.

One, the intrinsic/extrinsic distinction is not equivalent to the essential/accidental distinction (see Humberstone 1996, p.205). For any property, its being intrinsic or extrinsic depends on the nature of the property itself. By contrast, its being accidental or essential depends on the nature of its bearer. For instance, that triangles possess three-sidedness essentially tells us something about triangles, not something about three-sidedness. Plausibly, the same property can be both accidental

[6] Pereboom (2011, p.93) attributes this distinction to Kant, and notes that attention is called to it by Van Cleve (1988).

or essential depending on what bears it. Furthermore, the essential properties of a thing need not be intrinsic, and its accidental properties need not be extrinsic. There is no equivalence here.

Two, the intrinsic/extrinsic distinction is not equivalent to the categorical/dispositional distinction. Dispositions are propensities to manifest a certain property under certain circumstances. For instance, a vase's property of being fragile is (roughly) its propensity to break under the circumstance of being struck. By contrast, categorical properties are not bound to triggering events (see McKitrick 2003, p.351). Relations play an integral role in our understanding of dispositions but, as we will see in Section 2.3, it would be rash to claim that dispositions are extrinsic properties. Similarly, there are important links between being an intrinsic property and being a categorical property, but the two notions are not extensionally equivalent. For instance, the spatial relations of an object are categorical but extrinsic. Arguments akin to IA are often phrased in terms of our being ignorant of the categorical basis of physical dispositions. Stoljar, for example, characterises the Russellian position in terms of our conceptual ignorance of categorical properties (2006, Chapter 6). However, Pereboom (2011 pp.89-91) shows that it is an implicit notion of absolute intrinsicality that really drives those arguments, as we will see for ourselves in the next section.

Three, the distinction is between kinds of property, not kinds of predicate. It is tempting to say that monadic predicates correspond to intrinsic properties and polyadic predicates correspond to extrinsic properties. This would be a mistake since the same property can often be designated via both monadic and polyadic predicates, especially if we start introducing novel predicates. For instance, Socrates's height could be designated by a two-place predicate specifying the distance relation between his top and his bottom. Similarly, Socrates's property of being shorter than Simmias could be designated by a monadic predicate ascribed to the pair taken as a single complex object. We are pursuing a metaphysical distinction concerning properties, not a semantic distinction concerning predicates. Some reject such a distinction between predicates and properties, but that would be an error. Martin & Heil shed light on this:

> Some predicates designate properties, no doubt, and, in general, predicates hold true of objects in virtue of properties possessed by those objects. But it would be a mistake to imagine that every predicate, even every predicate that figures in a going empirical theory, designates a property. (1999, p.44)

With these terminological clarifications in place, we can now move on to evaluate the premises of IA.

SECTION 2

THE CASE FOR INSCRUTABILITY

2.1. THE RECEPTIVITY OF KNOWLEDGE

How do we gain epistemic access to the properties of external objects? The obvious answer is that we do so through perception, but that invites the further question of how perception reveals those properties. Shoemaker explains that '…properties reveal their presence in the actualizations of their causal potentialities, a special case of this being the perception of a property' (1997, p.242). Perceptual states are the effects of external objects on our senses, and are thus manifestations of dispositions belonging to those objects. It is highly plausible that we only gain epistemic access to physical objects through how they affect us because, as Kant puts it, '…how should our capacity for knowledge be awakened into action, if objects did not affect our senses, and partly of themselves produce representations…' (B1).[7] Kant calls this the *receptivity* of human knowledge (A26/B42). This claim - henceforth 'Receptivity' - is captured by IA1.

From how objects affect us we can infer a great deal about their relations to each other, including their causal relations. As Russell explains, '…in drawing inferences from percepts to their causes, we assume that the stimulus must possess whatever structure is possessed by the percept, though it may also have structural

[7] All Kant references are to the *Critique of Pure Reason* with page references for the A and/or B editions (1781/1787).

properties not possessed by the percept.' (1927, p.400) We have epistemic access to causal manifestations of physical objects, and to whatever can be inferred from those manifestations, but nothing more. Knowing objects through how they affect us allows us to build up a picture of the causal structure of the physical world – of a rich web of dispositions in a spatio-temporal framework.[8]

Our use of increasingly sophisticated measuring instruments provides a window on the world that overcomes the limitations of unassisted perception. This is consistent with Receptivity. When we use a Geiger counter, for instance, we are affected *indirectly* by radioactive particles. The particles affect our measuring instrument and the instrument affects our senses. Measuring instruments are only epistemically relevant insofar as we are affected by them and they are affected by whatever it is they detect. For instance, a Geiger counter only detects radioactive particles through their disposition to affect it. Epistemic contact always requires a chain of manifestations beginning with the object of knowledge and ending with the perceptual faculties of the subject. As our ways of accessing physical objects become more advanced, it remains the case that we only have epistemic access to them through the manifestation of their dispositions. As Blackburn puts it, '…science only finds dispositional properties, all the way down.' (1990, p.63)

Perceptual states may seem to reveal more than the dispositions of a stimulus. When an apple looks red, for instance, that redness appears to exceed any purely dispositional characterisation (see Blackburn 1990, p.65). To make sense of this, we must appeal to something like a secondary-property model of colours and their kin. Our perceptual experiences are indeed characterised by non-dispositional qualities. These qualities, however, are phenomenal properties of our *experience*, not properties of the *stimulus*. Phenomenal properties are absolutely intrinsic properties, but these properties do not belong to external physical objects. As such, our epistemic access to phenomenal properties is not a counter-example to inscrutability. Phenomenal qualities may reveal something about the stimuli that cause them; they reveal, for

[8] A complication I will not consider is that inscrutability may imply that we have no access to the intrinsic nature of time and space; that they are just the know-not-whats that cause our spatiotemporal experiences (Foster 2008, p.75). If this is true, it will have little bearing on IA, so there is no harm in putting it aside.

instance, the apple's *disposition* to generate in us experiences with red phenomenal qualities. They do not, however, disclose the non-dispositional features of the stimulus. Phenomenal qualities may *appear* to belong to stimuli rather than being caused by them but, on this account, that appearance should not be taken at face value. As such, it remains plausible that perception reveals only the dispositions of physical objects (see Foster 2008, p.51).

There are other apparent counter-examples to Receptivity. Extension, solidity and theoretical properties such as 'being an electron' do not appear to be purely dispositional, yet they are properties to which we have epistemic access. We will deal with these examples in due course but, in the meantime, it should be noted that Receptivity is a highly plausible claim about our epistemic access to external physical objects. As such, we can reasonably expect that any apparent counter-examples will not stand up to scrutiny.

2.2. THE RUSSELLIAN PICTURE

Receptivity – the first premise of IA – claims that we only have epistemic access to physical objects via how they affect us i.e. through the manifestations of their dispositions. This epistemic situation only entails an epistemic *limitation* if we have reason to believe that there is more to physical objects than their causal manifestations disclose. This is where IA2 and IA3 come in. Intuitively, there is more to objects than the causal relations in which they stand and their propensities to enter into different causal relations in different circumstances. We will put this intuition to the test in Section 3, but for the time being we will take it as given. The thought is that accessing an object through its manifestations tells us what it *does* but not what it *is*. When we know the causal *relations* in which physical objects stand we do not thereby know the nature of the *relata* themselves. Put another way, we can access the *structure* of the physical world but not the nature of the entities that *implement* that structure. As Russell states, we can '…infer a great deal as to the structure of the physical world, but not as to its intrinsic character' (1927, p.400).

What sense of 'intrinsic' drives this conclusion? The key claim is that relational properties presuppose a foundation of non-relational properties that characterise their relata. These non-relational properties must be *absolutely* intrinsic properties rather than merely comparatively intrinsic properties. The comparatively intrinsic properties of an object reduce to the extrinsic properties of that object's constituent parts. If there is a need for extrinsic properties to have a foundation of intrinsic properties, then an appeal to comparatively intrinsic properties merely pushes that need down a level. The extrinsic properties responsible for that comparatively intrinsic property would themselves need a foundation, and the sequence only ends when we have intrinsic properties that do not reduce to the purely extrinsic properties of their parts i.e. when we reach *absolutely* intrinsic properties.

Note, Receptivity does not render comparatively intrinsic properties epistemically inaccessible: an object's comparatively intrinsic properties can be inferred from a knowledge of the extrinsic properties of its parts. No such inference is available in the case of absolutely intrinsic properties. I take it that the structure of the physical world is the sum total of the extrinsic (and comparatively intrinsic) properties instantiated in the world. Sometimes philosophers have a more conservative sense of 'structure' in mind that excludes certain relational properties, but inscrutability is not committed to our ignorance of these relational properties.[9]

We can reinforce the thought that physical entities must have absolutely intrinsic properties, not just structural properties, by thinking about self-subsistence. Plausibly, physical objects are self-subsistent entities that are capable of independent existence (though objections to this view will be considered in the next section). They could even exist in a 'lonely' world in which there are no other physical entities with which to stand in relations. In other words, physical objects are *substances*. This has clear implications for the properties of physical objects. Langton, in an argument she attributes to Kant, explains these implications as follows:

[9] For instance, it could be held that we only know the *formal* structure of the physical world. We know that *some* objects stand in *some* relations in a certain logical form, but do not know what they are. Russell leans towards this extreme view, making him vulnerable to the 'Newman Objection' (Newman 1928, Demopoulos & Friedman 1989). This is the observation that any set of entities with the right cardinality can be construed as satisfying a formal description of the world, meaning that purely formal knowledge is almost entirely vacuous. Nothing in IA commits us to this extreme position: we can know what the relations are in which physical objects stand even if we cannot know their intrinsic nature.

> A substance is the kind of thing that can exist on its own: it can exist and be lonely. But nothing can exist without having properties. If a substance can exist on its own, it must have properties that are compatible with its existing on its own. If a substance can be lonely, it must have properties compatible with loneliness. So a substance must have intrinsic properties. (1998, p.19)

More specifically, physical substances must have *absolutely* intrinsic properties. Assuming that dispositions belong to substances, when we detect a disposition we can infer the existence of something with absolutely intrinsic properties.

The thought that we have no epistemic access to the intrinsic nature of objects appears in various forms throughout the history of philosophy. For instance, Berkeley writes:

> We see only the appearances, and not the real qualities of things. What may be the extension, figure or motion of anything really and absolutely, or in itself, it is impossible for us to know, but only the proportion or relation they bear to our senses. (quoted Lockwood 1989, p.155)

Similarly, Locke argues that we have no comprehension '…of the internal constitution and true nature of things, being destitute of faculties to attain it' (1690, II xxiii). Hume, in the same vein, holds that '…modern philosophy leaves us with no just or satisfactory idea of…matter.' (1739, 1.4.4) Kant argues that '[w]e have no insight whatsoever into the intrinsic nature of things.' (A277/B333) Indeed, his doctrine of the unknowability of things as they are in themselves can be understood as a variant of the inscrutability claim.[10]

There are also a number of more recent positions that support inscrutability. Feigl claims that physical entities are '..."triangulated" on the basis of various areas of observational (sensory) evidence…' but holds that '[w]hat these objects are acquaintancewise is left completely open...' (2002, p.71). Jackson explores what he calls a 'Kantian Physicalism' according to which we label entities by their relational properties but have no access to their 'intrinsic essences' (1998, p.23). Blackburn suggests that we have knowledge of dispositions located in space but not of the

[10] In making this observation I am not committing to Langton's (1998) reading of Kant, which seems to strip away the idealist components of his position. The observation is just that Kant, whatever else he says, uses the premise of Receptivity to reach the conclusion that we cannot know the absolutely intrinsic properties of objects.

properties that 'fill in' space (1990). Seager states '…we have, it seems, absolutely no knowledge of the intrinsic properties of matter which underwrite their causal relations.' (2006, p.135). Foster offers particularly powerful arguments for inscrutability, concluding that:

> ...in sharp contrast with the richness of our knowledge and potential knowledge of the spatio-temporal and functional character of the material realm, properties of intrinsic content are wholly beyond the reach of empirical discovery—or, at least, of discovery in a form that reveals what these properties are. (2008, p.47)

The claim of inscrutability has an impressive philosophical pedigree bridging a range of periods, areas of enquiry and philosophical persuasions. This does not show that the claim is *true*, but it does indicate that it should be taken seriously.

2.3. ARE DISPOSITIONS INTRINSIC PROPERTIES?

So far we have captured the thought that physical entities must have absolutely intrinsic properties. It is worth noting that a number of different positions can be taken on the *connection* between an object's intrinsic nature and its dispositions. Obviously an object's property of 'causing X' is an extrinsic property, depending as it does on X's presence. Consider the hammer's property of 'breaking the vase'. This cannot be an intrinsic property of the hammer since it clearly depends on something beyond itself, namely the vase and the circumstances that allowed the breakage (see Martin & Heil 1999, p.47). The interesting question is whether the hammer's *disposition* to break the vase is an intrinsic property. After all, the hammer's latent *potential* to break the vase does not depend on the existence of the vase, nor on the occurrence of the appropriate circumstances. The question is whether that disposition – that potential to enter into a certain causal relation – is routed in the nature of the hammer itself or has some source outside it.

There are three main positions that can be taken on the relationship between an object's intrinsic properties and its dispositions. First, an object's intrinsic properties could be taken to *necessitate* its dispositions. On this view, all duplicates of

an object have the same causal powers. Of course, whether that power is manifest depends on the possible world the object is in, but the potential itself is invariant. Objects would even have their dispositions in a lonely world. On this view, an object's causal powers are *grounded* in its intrinsic nature.

Second, an object's intrinsic properties could determine its dispositions *contingently*. On this view, the powers of an object arise from a combination of its intrinsic properties and the laws of nature. Duplicates of that object in worlds with different laws of nature would then have different causal powers. Though intrinsic properties play an important role here, it is not the case that an object's dispositions come out as intrinsic as they depend on laws that exist beyond the object itself. This position raises some serious problems. In particular, the notion of laws of nature being entities that somehow bestow causal powers on intrinsic properties is questionable. The first view offers a more plausible view of laws of nature: it is the causal powers of objects that are primary and the laws summarising those powers that are derivative, not the other way around.

Third, an object's intrinsic properties could be regarded as causally inert. On this view, objects have intrinsic properties and have dispositions, but there is no connection between them. Intrinsic properties do not *do* anything.[11] This position is implausible. Regarding its view of intrinsic properties, what does it mean for a property to be instantiated but to make no difference, actual or potential, to the world in which it is instantiated? As Heil suggests, '[p]roperties that make no difference to what their bearers do or would do are aberrations.' (2005, p.352) Regarding its view of dispositions, there is an issue of parsimony. On the first view dispositions are not onticly distinct from the intrinsic nature of objects. Why should we adopt a view that instead regards dispositions as basic properties when there is a more economical position available? Furthermore, it is hard to make sense of an object having dispositions that are not grounded in its intrinsic nature, but are instead instantiated as primitives.

I will not get side-lined by the debate surrounding the metaphysical status of dispositions, but the most plausible view is that an object's dispositions are

[11] Langton (1998) attributes this position to Kant, but it is not clear that Kant really thinks this.

necessitated by its intrinsic properties. I will continue on the assumption that this position is correct, though most of what I say about inscrutables could be adapted *mutatis mutandis* to match the second view. Adopting the third view would change things somewhat, but since this is a view that is not (and should not be) taken that seriously, there is no harm in putting it aside. The key conclusion here is that inscrutables are not causally inert.

Given that an object's dispositions are grounded in its intrinsic properties, must grounding properties be *absolutely* intrinsic or could they be *comparatively* intrinsic? Comparatively intrinsic properties may well confer dispositions. For instance, a vase has a disposition of fragility in virtue of its molecular structure. The vase has that molecular structure intrinsically: it is loneliness-compatible and duplicate-invariant. This property reduces to the vase's parts (its molecules) having certain extrinsic properties (their structure). This means that the vase's disposition of fragility is conferred by a *comparatively* intrinsic property. However, we can then ask where the molecules that constitute the vase get *their* dispositions from. To avoid a regress, it is plausible that the buck must eventually stop with dispositions that are conferred by *absolutely* intrinsic properties.

This plausible view of dispositions seems to threaten IA3. If dispositions are ultimately conferred by absolutely intrinsic properties, then we should have epistemic access to those properties through our receptive knowledge of an object's dispositions. There is a sense in which this is correct: when an object affects us, it does so in virtue of its absolutely intrinsic properties. As such, we can know those properties are present. To show what's wrong with this line of thought, we must deploy the distinction between transparent and opaque ways of knowing first mentioned in Chapter 4 (Section 2.1.1). It is one thing to know that a property has certain manifestations and quite another to know what that property *is*. Receptivity provides us with, at best, an opaque knowledge of absolutely intrinsic properties. We can designate such a property indirectly as the 'x' such that it confers given dispositions. But this falls short of a direct transparent mode of access. In the context of Kant, Allais explains:

> The claim is not that there are properties which are not powers - causally inert properties which have no implications for how objects interact with other things - but rather that describing a property as a power does not describe it as it is in itself. (2006, p.161)

Absolutely intrinsic properties exceed any purely relational/dispositional characterisation. To know such properties via their manifestations inevitably leaves out what those properties are like. One way of illustrating this is to consider two worlds W and W* occupied by entities that implement isomorphic causal structures but which differ with respect to the absolutely intrinsic nature of those entities. In light of Receptivity, those two worlds would be epistemically indistinguishable to us, thus we have no epistemic access to the intrinsic nature of physical entities.

It could be objected that it is *impossible* for two causally isomorphic worlds to differ with respect to the intrinsic nature of the entities implementing that structure. There are two responses to this objection. One, even if an object's intrinsic nature fixes its dispositions, it remains plausible that its dispositions do not fix its intrinsic nature. On this view all structural characteristics are multiply realisable, so a pair of entities could have the same dispositions but differ intrinsically. Two, even if all intrinsic differences entail some dispositional difference, a pair of entities could have the same *manifest* dispositions but differ in their latent dispositions. As such, the entities in W and W* could differ in their latent dispositions and so – unbeknownst to us – differ in their intrinsic nature.

The claim of inscrutability is not just that we do not *know* the intrinsic nature of actual physical objects, but that we have no *concepts* that characterise their intrinsic nature (Foster 2008, p.65). We can understand in abstraction that there is a difference between W and W*, but we lack the conceptual tools to specify (transparently) what that difference consists in. What justifies this claim of conceptual ignorance? In light of Receptivity, it is hard to see how we could form concepts of absolutely intrinsic properties. However, attempting to develop a robust and uncontentious model of our concept-forming procedures would be over-ambitious. A more practical route is to look at the kind of concepts we actually have, and see whether they offer a transparent conception of absolutely intrinsic physical properties.

2.4. EXTENSION AND SOLIDITY

It is tempting to claim that extension and solidity are absolutely intrinsic properties of which we have a transparent conception. Closer scrutiny, however, suggests that such a claim would be mistaken. Leibniz offers an illuminating account of extension, according to which it is a comparatively intrinsic property that can be resolved into the plurality, continuity and co-existence of its parts. As Pereboom explains, '[b]eing one of a collection of more than one thing, being continuous with other things, and coexisting with other things are all purely extrinsic properties of whatever has them.' (2011, p.93) This position is compatible with the infinite divisibility of parts: one extended part may be divisible into further extended parts, and so on *ad infinitum*, but it remains the case that a part's property of being extended is not absolutely intrinsic. On this view of extension, shapes come out as comparatively intrinsic. Being spherical, for instance, is simply a way of being extended.[12]

What about solidity? Solidity is plausibly an object's disposition to keep other objects out of the space that it occupies. It could be held that solidity is the absolutely intrinsic property *in virtue of which* an object has the disposition of impenetrability. However, in this case we would only have an opaque conception of solidity as the property responsible for that disposition.[13] We have no conception of solidity beyond its dispositional profile.

2.5. THEORETICAL TERMS

Do theoretical concepts, such as *being an electron*, characterise the intrinsic nature of physical entities? One might think that being an electron is something that stands apart from what an electron does. However, it is more plausible that '...the reference of the term 'electron' is fixed…by specifying the positions that electrons occupy in

[12] Leibniz proposes that extension is grounded in force. If he is right, this would not be a counter-example to inscrutability, as we only comprehend force opaquely through its causal manifestations.
[13] Pereboom offers an illuminating exposition of the concept of solidity with reference to Locke (2011, pp.97-100). Also see Foster (2008, p.59).

causal-structural networks' (Maxwell 2002, p.350).[14] Coleman helpfully fleshes out this claim:

> Physics tells us what electrons…are only by telling us how they interrelate with protons, forces and the like. Electrons are proton attractors, they are electron repulsors, they react to forces in such-and-such ways, have a mass of $9.10938188 \times 10^{-31}$ kilograms – which tells us about the kinds of displacements we can expect them to produce – the list continues.' (2008, p.87)

It could be held that to be an electron is to have a certain intrinsic nature – specifically, the intrinsic nature that the entities in fact performing the electron-role have. Even so, the concept 'electron' would only provide an *opaque* grasp of the intrinsic nature of those entities. This indirect designation of absolutely intrinsic properties provides no transparent characterisation of what those properties are like.

We do not need to go through each of our theoretical concepts before concluding that none of them offer a transparent characterisation of the absolutely intrinsic properties of physical entities? An examination of how theoretical concepts work indicates that it is *inevitable* that no theoretical concept is a counter-example to inscrutability. Blackburn explains that we can '…see the theoretical terms of science as defined functionally, in terms of their place in a network of laws' (1990, p.63). On this view, all theoretical terms work by specifying the status of entities in a causal system, so could not possibly characterise their intrinsic nature.[15] Lewis's (1970) account of theoretical terms, which builds on the ideas of Ramsey, sheds light on this claim. Sometimes theories introduce new terms into our vocabulary. Lewis's central idea is that any new term in a theory can be defined explicitly using the old vocabulary in that theory (see Cruse 2004, p.139). If we accept that our non-theoretical concepts cannot characterise the intrinsic nature of entities then, on this view, theoretical concepts will fare no better. Lewis emphasises that new theoretical terms refer to properties via their causal role, and that '[n]o amount of knowledge about what roles are occupied will tell us which properties occupy which roles.' (2009, p.204) This position on the

[14] Also see Ellis & Lierse (1994, p.32).

[15] The functional-role model of conceptual analysis discussed in Chapter 1 (Section 2.2.1) also complements this outlook.

limits of scientific representation, which Lewis dubs 'Ramseyan Humility', complements inscrutability.[16]

Staying on the topic of scientific representation, Epistemic Structural Realism (ESR) is another view that appears to complement inscrutability. However, I argue that ESR is a red herring. ESR was introduced by Worrall (1989) in response to a problem called the 'pessimistic meta-induction'. According to the pessimistic meta-induction '…we should infer that all theories scientists ever produce are false, on the inductive grounds that all past theories have been found to be false.' (Lipton 2000, p.182) Worrall attempts to get round this problem by distinguishing between the structural and the non-structural content of scientific theories. He suggests that the non-structural content of theories has usually turned out to be false, conceding that the non-structural content of our *current* theories should therefore not be trusted. However, Worrall suggests that the *structural* content of theories is generally retained when new theories replace old theories. For instance, Worrall claims that when Maxwell's electromagnetic field theory supplanted Fresnel's theory of elastic solid aether, it preserved the structure of the older theory but rejected its non-structural claims about the entities grounding that structure (1989, p.117). This continuity in structural content means that science's track-record presents no threat to the structural content of our current theories. In light of this, Worrall concludes that we are justified in believing what our best theories say about the structure of the physical world, but should not believe what they say about the intrinsic nature of entities occupying that structure.

What is the relationship between ESR and inscrutability? Like inscrutability, ESR claims that we are ignorant of the non-structural intrinsic nature of physical entities. Unlike inscrutability, it claims that we possess concepts for non-structural properties. For instance, according to Worrall the solid elastic aether theory makes substantive claims about the intrinsic nature of physical entities. He just holds that we are not justified in *believing* those claims. Is Worrall right that theories have non-structural content? No - it is plausible that the content Worrall regards as non-structural is, on closer examination, structural. 'Aether', for instance, is simply a notion of the 'X' that

[16] An interesting related view of theoretical concepts is presented by Rosenberg (2004).

does the things that aether does (Cruse 2004, p.138). As Papineau objects, the '…restriction of belief to structural claims is in fact no restriction at all' (1996, p.12). If we reject ESR, are we then lumbered with the pessimistic meta-induction? Again, no. Many other responses to this challenge are available. Lipton, for example, argues persuasively that the nature of scientific enquiry is such that we should *expect* a string of false theories before we start getting things right, thus blocking the inference from past falsity to present falsity (2000, p.202). Overall, ESR is not an ally of inscrutability.

SECTION 3

INSCRUTABILITY VS. PURE STRUCTURALISM

Pure Structuralism denies IA2 – the claim that physical entities have absolutely intrinsic properties. Pure Structuralism has been taken increasingly seriously, and presents an important threat to inscrutability. On this view, physical entities have no hidden aspect. It may well be that we have no concepts of absolutely intrinsic physical properties, but since physical entities do not *possess* such properties, this conceptual limitation does not entail ignorance of a type of actual physical truth. As Chakravartty explains, position aims to '…collapse the distinction between knowledge of structures and knowledge of natures.' (2004, p.152) In Section 3.1 I will outline Pure Structuralism. In 3.2 and 3.3 I will discuss the two main arguments for Pure Structuralism. In 3.4 I will conclude that Pure Structuralism is incoherent, and that IA2 is safe.

3.1. WHAT IS PURE STRUCTURALISM?

Pure Structuralism is the view that all physical entities ultimately have exclusively extrinsic properties. They may have *comparatively* intrinsic properties, since these are reducible to extrinsic properties. They do not, however, possess any absolutely intrinsic properties. The simple argument against Pure Structuralism is that extrinsic

properties must be possessed by self-subsistent objects and, as previously discussed, self-subsistence requires the possession of absolutely intrinsic properties. Some versions of Pure Structuralism deny the first of these claims and hold that the physical world is a network of extrinsic properties devoid of objects. Other versions deny the second claim and hold that extrinsic properties *do* presuppose objects, but those objects have an exclusively relational nature. On this view, objects are 'reconceptualised in a structuralist way' (Psillos 2006, p.561). The difference between these two positions is mainly terminological: both positions deny the existence of objects in the sense of substances with absolutely intrinsic natures. Pure Structuralists adopt '…eliminativism about self-subsistent individuals.' (Ladyman & Ross 2007, p.130)

Though it is intuitive that the concept of structure presupposes the existence of entities that exist independently of that structure, and so ultimately the existence of entities with non-structural properties, Pure Structuralism simply rejects that intuition. To cite this conceptual dependence as an objection to Pure Structuralism would thus be to beg the question (see Chakravartty 2003, p.872). Even if it is true that we cannot *think* of structures without a foundation of self-subsistent objects, the Pure Structuralist can dismiss this as a limitation of our cognitive capacities rather than a metaphysical insight (see Ladyman & Ross 2007). Overall, an argument against Pure Structuralism will have to deploy something more than the intuition that relations require self-subsistent relata.

Different versions of Pure Structuralism involve different notions of 'structure'. The most liberal form of Pure Structuralism is open to any extrinsic (or comparatively intrinsic property) in its ontology. If this kind of position can be defended, IA is in trouble. As such, there is no need for us to engage with the more conservative forms of Pure Structuralism according to which familiar extrinsic properties - such as spatial, temporal and causal relations - must be purged from our ontology along with absolutely intrinsic properties. Clearly, the threat to inscrutability does not rely on an

extreme formulation of Pure Structuralism, such as the view that the physical world is a purely logico-mathematical structure.[17]

The Pure Structuralist suggests that we revise our understanding of those entities that we take to be self-subsistent objects. When we sit on a chair, there is a locus of force that resists our body. There is a disposition of impenetrability in a spatio-temporally located area, but there are no absolutely intrinsic properties 'filling in' that area. There is simply the bare power itself. This power is object-like in that it remains constant through time, but it is not carried by any self-subsistent entity.[18] This power cannot be understood independently of the difference it makes to the wider world, just as a point on a graph is something that cannot be understood independently of its place in the graph as a whole.[19] Foster explains the view this encourages of the nature of particles:

> Instead of thinking of these particles as possessors of intrinsic content—for example, as items of some kind of space-occupying stuff—we should have to think of them simply as mobile items of causal power, with no further space occupant to form the vehicle of the power or cluster of powers involved, nor any non-functional properties on which the power or power cluster is nomologically grounded. (2008, p.66)

What motivates this kind of position? In the context of the philosophy of science, positions of this kind have been put forward as variants on ESR. We have seen how ESR attempts to evade the pessimistic meta-induction by claiming that we should only commit to the structural content of scientific theories. Some make the same move against the pessimistic meta-induction, but add the Pure Structuralist claim that physical entities do not have any non-structural properties.[20] I will dismiss this as an argument for Pure Structuralism for two reasons. First, as I have already suggested, the pessimistic meta-induction is a dubious challenge to scientific realism, so Pure Structuralism's promise of evading that challenge is of little importance. Second,

---

[17] Cao (2003) offers serious objections to the logico-mathematical version of Pure Structuralism, though Saunders (2003) responds that the positions Cao targets do not really adopt such an abstract view of structure.

[18] This view should not be confused with Dispositional Essentialism (Bigelow, Ellis and Lierse 1992) according to which an object's dispositions make it the kind of object it is. Dispositional Essentialism is compatible with the self-subsistence of objects.

[19] For more on this oft-cited graph analogy, see Ladyman (2007) and Bird (2007, p.534).

[20] This position is usually labelled 'Ontic Structural Realism' (e.g. Ladyman and Ross, 2007).

epistemic considerations about the track-record of science distract from the more important (and less contentious) epistemic considerations concerning the receptive nature of knowledge. The real reason for thinking our access to the physical is limited to the structural comes from Receptivity, not from the history of science.

I will consider two arguments for Pure Structuralism. The first is based on empirical considerations that are held to show that an ontology of self-subsistent objects is incompatible with certain scientific data. The second is based on methodological considerations that suggest Pure Structuralism should be our default position. On this view, positing absolutely intrinsic properties is compatible with the scientific data, but would be methodologically gratuitous.

3.2. THE EMPIRICAL ARGUMENT FOR PURE STRUCTURALISM

Some Pure Structuralists claim that empirical findings suggest that we should not adopt an ontology of self-subsistent individuals. Though object-hood is a perfectly useful notion when talking about tables and chairs, it should not be used in the context of fundamental physics. Consider, for instance, the question of whether quantum fields should be seen as a collection of particles or as a single entity with particle-like manifestations. Ladyman & Ross (2007) argue that the evidence 'underdetermines' the answer to this question. In some contexts, it looks like the field is simply an aggregate of particles. However, other findings suggest that the field is somehow primary and exceeds the features of the particles. As such, we can justify neither the view that particles are individuals, nor the view that the field is an individual.

Pure Structuralism can accommodate the scientific data here. It can hold that these different contexts encourage two illuminating but inconsistent object-based interpretations of a reality that is in fact purely structural. The world is neither one way nor the other *qua individual*, since it does not contain self-subsistent individuals. Notions like 'object' and 'individual' are merely useful tools. Some contexts encourage those tools to be deployed one way, other contexts encourage a different use, but in neither case do we capture the real nature of the phenomena in question. French &

Ladyman argue that '…the metaphysical packages of individuality and non-individuality would then be viewed in a similar way to that of particle and field in [Quantum Field Theory], namely as two different (metaphysical) representations of the same structure' (2003, p.37). But if there is no place for a notion of individuality in our metaphysics, there is no longer a place for the absolutely intrinsic properties of individuals.

This argument against self-subsistent individuals is unconvincing. It is not clear that there is indeterminacy between the particle-based and field-based positions. Cao, among others, argues that we have good reason to take fields as primary (2003, p.63). As such, fields are the self-subsistent individuals and bearers of absolutely intrinsic properties, not particles. This may be a strange discovery about the nature of fundamental entities, but that should not cast doubt on the applicability of a concept of self-subsistent objects. As Esfeld & Lam state, '…what is challenging about quantum physics is not that there are no objects, but that the properties of objects are remarkably different from the properties that classical physics considers.' (2008, p.34)

3.3. THE METHODOLOGICAL ARGUMENT FOR PURE STRUCTURALISM

The more interesting argument for Pure Structuralism concerns the application of methodological principles. The object-based view and Pure Structuralism offer competing ontologies. It is fair to conclude that empirical considerations alone will not show us that Pure Structuralism is true. Instead, the decision must be made on the basis of methodological considerations. Given a choice between the two positions, which is the default position? Pure Structuralists claim that the onus is on their opponents to show that an ontology free of absolutely intrinsic properties is not viable. Since such an ontology *is* viable, we should adopt Pure Structuralism. I distinguish between two related arguments that can be made in favour of Pure Structuralism being the default choice: parsimony and epistemic accessibility.

The first point in favour of Pure Structuralism is that given a choice between two viable positions we should generally advocate the one that is most economical ontologically. Following Occam's Razor, we should never posit more entities than are

needed to account for the evidence. Pure Structuralists claim that we can make sense of the world without positing absolutely intrinsic properties. If we *can* do without these extra properties then, applying Occam's Razor, we *should* do without them. The theory that physical entities have absolutely intrinsic properties is *underdetermined* by the evidence. In light of Receptivity, our evidence only ever warrants positing extrinsic physical properties. Positing absolutely intrinsic properties as well would thus be methodologically unsound. One might try to reject Pure Structuralism on the basis that we *do* have epistemic access to absolutely intrinsic properties after all, but this is not an option for the inscrutability view. The conclusion of IA is precisely that we lack epistemic access to the intrinsic nature of physical entities.

The second consideration in favour of Pure Structuralism is that it is methodologically unsound to posit *unknowable* properties. Pure Structuralism and the inscrutability position agree that our epistemic access to the physical world is limited to its structure. In claiming that absolutely intrinsic properties are unknown, inscrutability is faced with a dilemma. On the one hand, it could posit inscrutables despite our lack of knowledge of them. But how could we possibly justify positing inscrutables if we cannot know that they exist? On the other hand, it could claim that we do have knowledge of inscrutables after all. But then the position becomes incoherent, claiming that we have knowledge of unknown properties.[21] Pure Structuralism does not posit unknown properties, so avoids this methodological issue. Another way of making this point is to consider what *purpose* the notion of inscrutables could possibly play in our intellectual practice. If there is no use in talking about inscrutables, surely we should prefer a position that excludes them from its ontology? Ladyman & Ross put this in stronger terms, arguing that '…no hypothesis that the approximately consensual current scientific picture declares to be beyond our capacity to investigate should be taken seriously' (2007, p.29).[22]

[21] This is a familiar objection to Kant's doctrine of the unknowability of things-in-themselves.
[22] Ladyman & Ross go as far as adopting a kind of neo-verificationism (2007, p.130). However, the point under discussion can be made without committing to this bold position.

3.4. THE INCOHERENCE OF PURE STRUCTURALISM

The considerations above plausibly show that *if* Pure Structuralism is a viable ontology *then* we should prefer it to a position that posits absolutely intrinsic properties. However, a strong case can made against the antecedent of this conditional. Pure Structuralism cannot be true. The thought that *relations require relata* can be reinforced in a way that casts serious doubt on the coherence of Pure Structuralism. Though Pure Structuralism denies the existence of self-subsistent objects, it affirms the existence of distinguishable extrinsic properties. The physical world-structure must consist in a plurality of extrinsic properties. The challenge to Pure Structuralism is this: how are these extrinsic properties to be individuated? What makes one instantiation of a spatio-temporal or causal property distinct from any other? Lowe elucidates this challenge as follows:

> The problem ... is that *no property can get its identity fixed*, because each property owes its identity to another, which, in turn owes its identity to another – and so on, in a way that, very plausibly, generates either a vicious infinite regress or a vicious circle. (quoted Bird 2007, p.523)

According to Pure Structuralism, the identity of a property is determined by its relations. But such relations presuppose the existence of properties that *stand* in those relations. Ladyman paraphrases this objection: 'Relations presuppose numerical diversity and so cannot account for it.' (2007, p.23) Interestingly, this objection to Pure Structuralism is an application of the motto that *relations require relata*, but avoids a question-begging presupposition that the physical world contains self-subsistent objects.

This problem of individuation becomes more obvious in the context of the Pure Structuralist's view of causal powers. Keith Campbell asks '[w]hen one point moves another, all that has been shifted is a power to shift powers to shift.... But powers to shift *what*?' (quoted Bird 2007, p.520). Similarly, Ellis states:

> If all of the properties and relations that are supposed to be real are causal powers, then their effects can only be characterized by their causal powers, and so on. So causal powers are never manifested. They

> just produce other causal powers in endless sequence. (quoted Bird 2007, p.520)

In order to make sense of powers, we need entities with absolutely intrinsic properties that can stand in causal relations, but which are not reducible to such relations. As Ellis & Lierse argue, '[r]eal dispositions involve real changes to the object in question. For example, solubility is a real disposition, for a soluble substance undergoes a genuine change when the disposition is manifested.' (1994, p.36)[23]

Bird responds to this line of attack by considering the identity of points on a graph, arguing that the identity of those points depends on the graph as a whole and cannot be discerned in isolation from it. Extrinsic properties, he claims, similarly rely upon the world-structure as a whole.[24] This line of thought is unconvincing. The identity of mathematical objects may well be relational, but *concrete* objects are a different matter. After all, to be concrete is essentially to be capable of existing apart from other things (Seager 2006, p.142). In a similar vein, Strawson holds that '…there is nothing more to a thing's being than its intrinsic, non-relational propertiedness.' (2006, p.28) In other words, abstract entities might have exclusively relational properties, but concrete entities cannot. We may well be able to *define* concrete entities in terms of their relations, but it does not follow that those entities could *consist* in nothing but relations. As Russell himself observes, defining something in terms of its relations '…always indicates some class of entities having…a genuine nature of their own.' (quoted Ladyman 2007, p.31) Van Fraassen captures this objection to Pure Structuralism succinctly: '*Structure of nothing is nothing*…' (2007, p.60). The difference between a real concrete structure and an abstract structure is that real structures have some *non-structural* foundation. As such, Pure Structuralism does not offer a viable picture of the *real* world.

If a Pure Structuralist ontology is impossible, how exactly should we respond to the arguments in favour of Pure Structuralism? Regarding any putative empirical evidence in favour of Pure Structuralism, we can simply respond that there can never

---

[23] A possible view is that all powers are ultimately powers to effect spacetime. Spatiotemporal properties would be non-power properties that halt the problematic regress, but are not absolutely intrinsic properties. However, this position is successfully undermined by Foster (2008, pp.69-72).

[24] Also see Holton (1999).

be empirical evidence for an impossible ontology. Regarding parsimony, given a choice between two possible positions we should indeed adopt the more economical position. However, Pure Structuralism is *not* a possible position, so its apparent parsimony is irrelevant.[25] To paraphrase Einstein, we should make our ontology as simple as possible, but not one bit simpler. Positing absolutely intrinsic properties is not underdetermined by the evidence. We can verify the presence of certain causal properties and, since causal properties cannot exist without absolutely intrinsic properties, we can infer the existence of absolutely intrinsic properties. As such, our evidence reveals – albeit indirectly – the presence of absolutely intrinsic properties. Of course, this inference from observations to the existence of inscrutables involves a significant element of conceptual analysis. Some proponents of Pure Structuralism '…deny that a priori inquiry can reveal what is metaphysically possible' (Ladyman & Ross 2007, p.16), so will not be convinced by this inference. However, this view of a priori inquiry is dubious, and it is already a commitment of this thesis that conceptual analysis *can* reveal modal truths.

Regarding the unknowability consideration, the inscrutability view should concede that positing totally unknowable properties is unsound. As Cao puts it, '…if something is cognitively "utterly inaccessible" in principle, then there is no point in talking about its empirical existence.' (2003, p.65) It should also be conceded that '[i]f it is claimed that there is something that exists but that we cannot know, we need an argument why we should accept that there is any such thing' (Esfeld & Lam 2008 p.29). However, if Pure Structuralism is impossible, these principles can be satisfied. We have evidence of absolutely intrinsic properties every time we are affected by an object. The existence of inscrutables is not a fanciful speculation; it is something we can infer through a combination of empirical evidence and a priori reflection. As such, we *do* know that there are absolutely intrinsic properties – we just don't have any conception of what they are.[26] Maxwell explains that:

---

[25] Furthermore, Chakravartty (2003) makes the interesting point that Pure Structuralism is not actually more economical than its competitors. Pure Structuralism is forced to posit clusters of primitive dispositions that could be better accounted for as the manifestation of a smaller set of fundamental intrinsic property.

[26] Langton (1998, p.13) uses an argument along these lines to defend Kant against the threat that the doctrine of the unknowability of things-in-themselves is self-defeating. In response, Cowling (2010) has

> …science *does* assert the *existence* of instances of a variety of intrinsic properties; moreover, it provides information about the various causal-structural roles that such instances play. However, it *does* leave us completely ignorant as to *what* these intrinsic properties *are*. (2002, p.350).

Talking about such properties may be of no use to scientific inquiry but, as our discussion shows, we may nevertheless have reason to posit their existence. In summation, Pure Structuralism does not present a serious threat to the Inscrutability Argument.

SECTION 4

INSCRUTABLES AND CONSCIOUSNESS

At the beginning of this chapter I explained that the Russellian Ignorance Hypothesis (RIH) consists in two claims: that we have no conception of the (absolutely) intrinsic properties of physical entities, and that those properties are integral to the explanation of consciousness. We have come far enough to conclude that the first of these claims is plausibly true. Now we must return to the matter of consciousness to assess whether an equally strong case can be made for the second claim. The fundamental thought is that the intrinsicality of inscrutables makes them plausible candidates for the explanation of intrinsic phenomenal properties. Though there is something valuable in this line of thought, we will see that it faces severe limitations.

4.1. RIH AND TYPE-F MONISM

The ignorance hypothesis asserts that we are conceptually ignorant of properties integral to the physical explanation of consciousness. The Russellian version of the

---

objected that 'categorial properties' such as 'having intrinsic properties' are intrinsic properties. As such, knowledge that physical objects have such properties is itself a counter-example to inscrutability. I respond (McClelland, 2012) with three objections to Cowling's argument, the most important of which is that categorial properties are not the kind of property knowledge of which is excluded by inscrutability.

ignorance hypothesis makes the more specific claim that those unknown properties are inscrutables. This qualifies RIH as a member of the 'Type-F Monist' family of positions on consciousness. Chalmers introduces this label as follows: 'Type-F Monism is the view that consciousness is constituted by the intrinsic properties of fundamental physical entities: that is, by the categorical bases of fundamental physical dispositions.' (2002, p.265). Though RIH plausibly qualifies as a form of Type-F Monism, it diverges significantly from other Type-F Monist positions. There are three possible positions that Type-F Monists can take on the relationship between fundamental intrinsic properties and phenomenal properties:

> 1) The fundamental intrinsic properties of physical entities are phenomenal properties that combine to form our phenomenal states.
>
> 2) The fundamental intrinsic properties of physical entities are qualities of which we are directly aware in consciousness, but which exist independently of such awareness.
>
> 3) The fundamental intrinsic properties of physical entities are robustly non-phenomenal properties and are integral to the physical explanation of phenomenal states.

According to '1' the intrinsic nature of fundamental physical entities is inherently experiential. Since fundamental physical entities are ubiquitous, this position is committed to *panphenomenalism*. Some proponents of this stance emphasise that fundamental physical entities have only a highly attenuated degree of consciousness. However, admitting *any* degree of phenomenal awareness puts you in category 1. According to '2', consciousness need not be ubiquitous, but the qualities of which we are aware in conscious experience *are* ubiquitous. We can call this position *panqualitativism*. A number of different views can be taken on which qualities are instantiated fundamentally and which (if any) are derivative.[27] Different positions can also be taken on the explanation of our awareness of these qualities. I will not discuss these variations here.

According to both 1 and 2, we *do* have concepts of the intrinsic nature of fundamental physical entities: our *phenomenal* concepts characterise the absolutely

[27] A terminological clarification is needed here: I will call a quality a *phenomenal* quality only when it is a property of a phenomenal state. One and the same quality can be both non-phenomenal or phenomenal depending on whether it is being experienced.

intrinsic properties of physical entities, or at least give us some grasp of those properties.[28] As such, neither 1 nor 2 are consistent with the ignorance hypothesis. Proponents of these positions do not respond to the Problem of Consciousness with anything like EV. Consequently, RIH must fall under category 3. This is the view that though inscrutables are integral to the explanation of the phenomenal, they are not themselves phenomenal properties or qualities i.e. they are robustly non-phenomenal.

Though RIH must not be conflated with other forms of Type-F Monism, it obviously has a lot in common with these positions. Type-F Monism has an interesting history, though it is not always clear which of the three categories a proposal falls under. Locke suggests that, in light of our ignorance of the intrinsic nature of physical objects, we cannot conclude that mind and matter are incompatible (see Megill, 2005). Similarly, Kant argues:

> …if we compare the thinking 'I' not with matter but with the intelligible which lies at the ground of the external appearance we call matter, we cannot say that the soul is intrinsically any different from it, since we know nothing at all of it. (A360)

Though Kant resists this mentalistic route, he does interpret Leibniz as adopting such a view (A274/B330). Moving forward a few centuries, thinkers such as Clifford adopted the explicitly panphenomenalist view that '…the reality external to our minds which is represented in our minds as matter, is in itself mind-stuff.' (1878, p.67)

Russell took the view that '…the ultimate constituents of matter are not atoms or electrons, but sensations…' (1921, Lecture VI) It is plausible that Russell's notion of 'sensation' is roughly equivalent to our notion of 'qualitative character'. As such, he seems to fall into category 2, holding that physical entities have a qualitative nature that is not inherently experiential, but with which we are directly acquainted in experience. As I have already argued, this is not the route that RIH takes. Though RIH is clearly 'Russellian' in spirit, it is not strictly Russell's own position. Feigl takes a similar view to Russell, holding that '…I am directly acquainted with the qualia of my own immediate experience, I happen to know (by acquaintance) what the

[28] I take it that any grasp we have on unexperienced qualitative redness is based on our concept of *experienced* phenomenal redness.

neurophysiologist refers to when he talks about certain configurational aspects of my cerebral processes.' (2002, p.71). Lockwood adopts a related position:

> Consciousness…provides us with a kind of 'window' on to our brains, making possible a transparent grasp of a tiny corner of a material reality that is in general opaque to us, knowable only at one remove. The qualities of which we are immediately aware, in consciousness, precisely *are* some at least of the intrinsic qualities of the states and processes that go to make up the material world - more specifically, states and processes within our own brains. (1989, p.159)

Similarly, Maxwell holds that '[i]f we recognize that C-fibre activity is a complex causal network…the way is left open for the neurophysiologist to theorize that some of the events in the network *just are pains* (in all of their qualitative, experiential, mentalistic richness).' (2002, p.347). Other proponents of category 2 positions include Unger (1998), Banks (2010) and Coleman (2006/2008).

The most prominent category 1 view is Strawson's panphenomenalist position (1994/2006). We have mentioned his view already and will consider it again shortly. Chalmers (1996) and Seager (1995, 2006) sympathise with the category 1 stance. Rosenberg (2004) offers a position that builds the experiential nature of fundamental intrinsic properties into a distinctive metaphysics of causation.[29] Category 3 versions of Type-F Monism have received less attention. As discussed, Stoljar (2001) originally adopted such a position and still takes it seriously in his reformed work (2006). Holman (2008), Montero (2010) and Pereboom (2011, p.197) all show sympathy with this kind of view.

Not all versions of the ignorance hypothesis adopt Type-F Monism and not all versions of Type-F Monism advocate the ignorance hypothesis. RIH lies in the intersection of these two groups. Though RIH must be carefully distinguished from other Type-F Monist positions, in Chapter 6 we will see that it is has something to learn from these existing proposals.

[29] Rosenberg's views on causation are criticised persuasively by McKitrick (2006).

4.2. RIH AND THE –TRINSICALITY GAP

The Relevance Condition on EV states that we should only adopt the ignorance hypothesis if we have reason to believe that unknown physical properties could evade the a priori obstacles to closing the epistemic gap. These obstacles are the –trinsicality and –tivity gaps, each of which indicate that *no* physical property could allow the epistemic gap to be closed. I will discuss how RIH addresses the –trinsicality gap here, and postpone discussion of the –tivity gap until Section 4.4.

4.2.1. Inscrutables and Phenomenal Qualities

Our discussion of the Inscrutability Argument has provided us with conceptual tools that can shed light on the –trinsicality gap. We can refine the four step argument for the –trinsicality gap from Chapter 1 (Section 3.2.2) as follows:

> TRIN1') All physical properties are structural properties i.e. extrinsic properties or comparatively intrinsic properties.
>
> TRIN2') All phenomenal states involve the instantiation of non-structural properties i.e. absolutely intrinsic properties.[30]
>
> TRIN3') There can be no epistemic entailment from extrinsic properties or comparatively intrinsic properties to absolutely intrinsic properties.
>
> TRIN4') Therefore, there can be no epistemic entailment from the physical to the phenomenal.

Clearly, RIH denies TRIN1'. It is not the case that all physical properties are structural properties. Some physical properties are *absolutely intrinsic* properties, though we have no concepts for them. Furthermore, RIH can account for the *prima facie* plausibility of TRIN1'. At present, we can only characterise physical entities in terms of their structural properties. We even have reason to believe that this a permanent feature of our epistemic situation. As such, it is natural for us to conclude that all physical properties are structural, but RIH seeks to cast doubt on this conclusion. On this view, the apparent plausibility of the –trinsicality gap is symptomatic of our limited conception of the physical.

[30] Note, the non-structural properties of phenomenal states are its *phenomenal qualities*. I will discuss the view that the *subjectivity* of phenomenal states is also non-structural in the next sub-section.

Can RIH defensibly maintain that inscrutables are robustly non-phenomenal? Since being an intrinsic property is not a sufficient condition of being a phenomenal property, or of being qualitative, it is permissible to claim that inscrutables are non-phenomenal. But if inscrutables are non-phenomenal, is it plausible that they are responsible for the intrinsic qualities of our conscious states? For instance, how could an intrinsic property that is not itself a property of qualitative redness be responsible for the red-quality of a conscious state? If the fundamental intrinsic properties of physical entities were straightforwardly qualitative, as they can be on the panphenomenal or panqualitative accounts, this problem would not arise. But RIH must hold that inscrutables are *not* themselves phenomenal or qualitative, yet are nevertheless integral to their explanation.

RIH can be defended against this kind of worry. To claim that phenomenal qualities cannot be accounted for by anything more basic than themselves is to adopt Primitivism about phenomenal qualities. As such, it begs the question against RIH. RIH claims that the intuition that phenomenal qualities are inexplicable in more basic terms is symptomatic of our limited conception of the physical. Because of our ignorance, RIH cannot *demonstrate* that phenomenal qualities are explicable in terms of inscrutables, but the onus is on the Primitivist to show that such an explanation is impossible. The Primitivist could argue that inscrutables are the *wrong kind* of property to explain phenomenal qualities, but how would such an argument go?[31] This is normally where the –trinsicality gap would come in, but this clearly has no force against RIH.

An attractive feature of the way RIH undermines the –trinsicality gap is that it respects TRIN 3'. The conceptual insight that there can be no epistemic entailment from the structural to the non-structural is all the more compelling when rephrased in terms of there being no entailment from extrinsic properties to absolutely intrinsic properties. This is a difficult claim to deny, and it would be ill-advised for EV to take the route of maintaining TRIN 1' and instead denying TRIN 3'. Contra Stoljar, it is deeply implausible that there could be unknown extrinsic physical properties that bridge the apparent conceptual gap between the extrinsic and the absolutely intrinsic. In fact, the

[31] This question will be considered further in Chapter 6 (Section 2.3).

very concept of absolute intrinsicality makes such a bridge impossible. This is important to our evaluation of EV: EV is defensible only if it undermines the –trinsicality gap, and EV can only undermine the –trinsicality gap by holding that the properties of which we are ignorant are absolutely intrinsic properties. Therefore the plausibility of EV depends on the plausibility of RIH.

4.2.2. Is Subjectivity Non-structural?

Now that we have the refined formulation of the –trinsicality gap offered by TRIN', a brief tangent is in order. In Chapter 1 (Section 3.3.1) I explained that the –trinsicality gap pertains to the qualitative character of phenomenal states rather than to their subjectivity (that is, to their being phenomenal states at all). I also explained that some take a different view on this and claim that there is a conceptual gap between the structural and the subjective (e.g. Chalmers 1996). According to this kind of position, it would be misleading to claim, as I have, that the –trinsicality gap pertains only to the *qualitative character* of conscious states. We are now in a better position to understand that view and explain why it is mistaken. There are two senses in which subjectivity may indeed be regarded as 'non-structural', but neither presents a serious obstacle to Physicalism.

First, we have mentioned that on some views physical facts are 'structural' in a conservative sense; that they are purely abstract and formal. According to this kind of position, there is nothing more to the physical world than the mathematical structures captured by physics. Plausibly, such *formal* facts can only entail *more* formal facts, but subjective awareness is not an abstract formal property of a state. It is true that subjectivity is a substantive non-formal characteristic. However, the problem with formulating the –trinsicality gap in this way is that it is *im*plausible that all physical facts are 'structural' in this conservative sense. Though many Physicalist ontologies reject inscrutables, they generally admit substantive relational properties into their ontology. As such, this line of thought does not present any serious obstacle to Physicalism, so is not a serious pillar on which to rest the case for Primitivism.

Second, there is a further sense in which subjectivity is non-structural. Seager, for instance, claims that '[t]he realization that states of consciousness are intrinsic properties is of great significance.' (2006, p.136) Applying our two diagnostic tests of intrinsicality, he goes on to suggest that '…my duplicates — even if the only thing in the world — would share all my states of consciousness.' (2006, p.136). However, though Seager and others are plausibly correct that properties such as 'being in a subjective state' are intrinsic to us, they would be wrong to make the bolder claim that they are *absolutely* intrinsic. Why should we rule out consciousness being a *comparatively* intrinsic property of an individual? Phenomenal qualities cannot be reduced to the extrinsic properties of their parts, but it is not immediately implausible to claim that the property of 'being in a subjective state' is reducible in this way. In this sense, subjectivity might be a *structural* property, so the (apparently) structural nature of physical facts presents no special obstacle to explaining subjectivity in physical terms. Comparatively intrinsic properties are derivable from extrinsic properties, so even if all basic physical properties are extrinsic properties there is no a priori obstacle to the entailment. The *objectivity* of those physical properties may present a problem, but that is a different matter. Overall, we can maintain the conclusion that the apparently structural nature of physical facts presents an obstacle to the physical explanation of the *qualitative character* of phenomenal states, but not to their *subjectivity*.

To conclude, RIH evades the first a priori obstacle to the explicability of consciousness in physical terms. As such, it goes at least half way to satisfying the Relevance Condition.

4.3. RIH AND THE INTEGRATION CONDITION

The ignorance hypothesis is only plausible if it can be integrated with what we already know about the physical world. There are two layers to the satisfaction of the integration condition: finding a blind-spot in our current conception of the physical, and showing that the properties occupying that blind-spot are suited to performing the

proposed explanatory role. I will consider how RIH fares with respect to these two layers in turn.

4.3.1. The Epistemic Status of Inscrutables

In Chapter 4 (Section 3.3.2) I established a challenge for any version of the ignorance hypothesis. We need positive evidence that the proposed unknown properties are actual. We also need an account of why it is we are conceptually ignorant of the proposed properties, and how we have managed to get by despite our ignorance. These criteria seem to pull in opposite directions – positing the unknown properties seems to require them to be manifest to us, but our ignorance of them seems to require them not to be. RIH manages to overcome this worry.

Despite our conceptual ignorance of inscrutables, we must posit their existence. A critic cannot hold that though we have no concepts of absolutely intrinsic properties, such properties are not actually instantiated. This would be to adopt an incoherent Pure Structuralist stance. We know that physical entities must have absolutely intrinsic properties, but also know that we have no concepts of what those properties are. Adopting Donald Rumsfeld's infamous phrase, we might call inscrutables a *known unknown*.

If inscrutables play an ineliminable role in the physical world, how have our explanatory practices been so successful despite our ignorance? Why does it appear to thinkers such as Prinz (2003) that the brain has no hidden properties? Foster provides an answer to these questions:

> …the limitation on what physical science can reveal…is not perceived as a practical limitation from the viewpoint of the scientist. He never finds himself wanting to evaluate hypotheses about the nature of particle content, since the possibilities for content are not scientifically specifiable. The point where the nature of the physical situation falls beyond the scope of empirical tests is the point where he runs out of vocabulary with which to formulate the options, and concepts by which to conceive of them. (2008, p.65)

Inscrutables are not the kind of property that we need to worry about when trying to explain phenomena (with the possible exception of consciousness). We want causal

explanations for things, and such explanations can be provided in terms of dispositions and the relations in which entities stand without having to specify the intrinsic ground of those dispositions. Worlds can differ with respect to the intrinsic properties of physical entities but, for almost all of our explanatory projects, this is not a *difference that makes a difference*. So long as our conception of the physical world's *structure* is adequate, we will not notice our conceptual ignorance. Nevertheless, philosophical reflection suggests that inscrutables still have an indispensable part to play in our ontology.

It is a significant virtue of RIH that it evades this epistemic quandary with such ease. It is hard to imagine other versions of the ignorance hypothesis that could offer compelling arguments for the existence of a conceptual blind-spot whilst also explaining how that blind-spot could generally go unnoticed. Though it would be rash to conclude that RIH is the *only possible way* of meeting this challenge, this does lend further support to the claim that RIH is EV's best hope.

On the topic of the epistemic status of inscrutables, it is worth noting that RIH indicates that our ignorance is *permanent*. I have already argued that whether or not our ignorance is chronic has no bearing on the evaluation of EV one way or the other. Nevertheless, it is worth recognising that RIH entails that the explanation of consciousness is permanently beyond our grasp. Montero adopts a more optimistic view of the Russellian approach:

> …given that physics has changed in ways that would have been inconceivable to earlier generations, it seems we should leave open the possibility that physics could, someday in the unforeseeable future, explain both structural [and] non-structural features of the world.' (2010, p.79)

In other words, Montero thinks it is premature to conclude that science will never provide us with a conception of inscrutables. This argument, however, seems to miss the point. It is not a contingent feature of science that it only describes structural properties – rather, the receptive nature of knowledge plausibly makes this limitation inevitable. Though some may find this commitment of RIH unpalatable, it should not be regarded as a defect of the position.

4.3.2. The Suitability of the Blind-Spot

RIH designates a specific blind-spot in our current conception of the physical. Given what we already know about the place of consciousness in the physical world, is that blind-spot a suitable home for properties that are integral to the explanation of consciousness? In connection with this question I introduced three criteria (Chapter 3, Section 3.3.1). RIH meets all of them.

First, what we know about the occupants of the conceptual blind-spot must tally with our conclusions about what kind of physical property can satisfy the Relevance Condition. So far, in response to the Relevance Condition I have argued that the unknown physical properties are absolutely intrinsic properties, and that their being so is what undermines the –trinsicality gap. The conceptual blind-spot revealed by the Inscrutability Argument *must* be occupied by absolutely intrinsic properties. As such, there can be no doubt that the proposed conceptual blind-spot is a suitable home for physical properties that evade the –trinsicality gap.

Second, we must be able to integrate RIH with what we already know about the physical basis of consciousness. For instance, we have reason to believe that the brain is the seat of consciousness. This suggests that if inscrutables are to play a role in the explanation of consciousness, they should be located in the brain. Since inscrutables are ubiquitous, they will indeed be instantiated in the brain. In fact, if *any* physical state is discovered to be a correlate of consciousness, that state will inevitably be grounded in inscrutables. Consequently, there is no risk of inscrutables being unsuitably located to play a role in the explanation of consciousness.

Third, RIH must integrate with what we know about the physical efficacy of conscious states. If inscrutables are integral to the explanation of consciousness, then the causal status of inscrutables will place restraints on the causal status of consciousness. The key point here is that inscrutables fall *within* the causally closed system of the physical. If they stood outside that system, RIH might have had trouble respecting the physical efficacy of conscious states. Interestingly, inscrutables are not extra components in the causally closed system, but rather the ground of components with which we are already familiar. We represent physical causes and effects

structurally, but now have reason to believe that these causal processes are founded on the intrinsic nature of physical entities.[32]

There is one remaining consideration that should be addressed in the context of the Integration Condition. Given what we understand about the physical world, is it acceptable to claim that inscrutables qualify as *physical* properties? EV is meant to defend Physicalism, so if inscrutables turn out to be non-physical, RIH fails. According to some strict standards of physicality, inscrutables are not physical because they do not figure in physical theory. As previously discussed though, this is not the sense of 'physical' relevant to the Problem of Consciousness. Inscrutables satisfy our two conditions of physicality: being non-phenomenal, and falling within the causally closed system of the physical. Of course, some Type-F Monist positions claim that the intrinsic properties of physical entities are *phenomenal*. In our terms, such positions are not Physicalist. However, we have been explicit that RIH is committed to inscrutables being non-phenomenal. It could be argued that we are in no position to *rule out* the suggestion that the intrinsic properties of physical entities are themselves phenomenal (see Megill, 2005), but this would be asking too much of RIH. The burden of proof lies with the anti-Physicalist, so RIH only needs to show that a viable Physicalist account of consciousness is available, not that all anti-Physicalist positions are demonstrably false.

Overall, RIH has no trouble satisfying the Integration Condition. Furthermore, any alternative version of the ignorance hypothesis would have its work cut out meeting the condition so successfully. Again, it seems likely that RIH is the best hope for EV.

4.4. RIH AND THE –TIVITY GAP

RIH has gone a long way towards satisfying both conditions on EV, and so towards offering a viable defence of Physicalism. However, this is as far as it goes. In order to satisfy the Relevance Condition, RIH must undermine both the –trinsicality and –tivity

[32] I will discuss concerns surrounding the causal status of inscrutables further in Chapter 6 (Section 2.5).

gaps. Though it succeeds with respect to the former, I will argue that it fails with respect to the latter.

4.4.1. The Objectivity of Inscrutables

The –tivity gap is based on three claims: that physical states are always objective, that phenomenal states are always subjective and that there is no epistemic entailment from the objective to the subjective. As we discussed in Chapter 3 (Section 3.2.2), there are three different ways in which unconceived physical properties might close this gap. The unknown properties may be a kind of objective property, knowledge of which would reveal an entailment from the objective to the subjective. Alternatively, the unknown properties may themselves be subjective, rendering the gap between the objective and the subjective irrelevant. Finally, the unknown properties might fall into a third category between the objective and the subjective, and be suited to explaining subjective states.

Since inscrutables are objective properties, RIH is committed to the first route. The instantiation of inscrutables does not, in and of itself, mean the instantiation of subjective awareness. Why, then, should we believe that these objective properties are any more suited to the explanation of subjectivity than familiar objective properties are? There is nothing special about inscrutables that would cast doubt on the conclusion that they leave the gap untouched. It could be held that since we do not have concepts of inscrutables, we cannot *rule out* their being relevant to the explanation of subjectivity. But, as we concluded in Chapter 3, the burden of proof lies with EV to provide *positive evidence* for their relevance. Sheer optimism is not enough.

It could be held that the reason Physicalists have difficulty accounting for subjectivity is that it is an intrinsic property, so cannot be explained in structural terms. On this view, the *intrinsicality* of inscrutables might make them good candidates for undermining that threat. However, as we have discussed, it would be a mistake to formulate the threat to Physicalism in this way. It is the *objectivity* of physical properties that underwrites the –tivity gap, and inscrutables do nothing to cast doubt on that gap.

4.4.2. Alternative Strategies

Maybe the best way to undermine the –tivity gap is to deny that inscrutables are objective. Perhaps, as Strawson argues, the intrinsic nature of fundamental physical entities is inherently subjectivity-involving. To say that inscrutables are inherently subjectivity-involving is to say that they are *phenomenal* properties: unlike intrinsicality, subjectivity is a sufficient condition of phenomenality. On this view, there is *something it's like to be* an entity with absolutely intrinsic properties. Physical structures are grounded on experiential properties. There are a number of problems with this position.

Claiming that inscrutables are phenomenal properties is at odds with the central claims of EV. The purpose of EV is to undermine the Primitivist position that phenomenal properties are fundamental. After all, EV is meant to defend Physicalism, and Physicalism requires all fundamental properties to be non-phenomenal properties. The panphenomenalist route concedes to Primitivism that there are basic phenomenal properties. It merely supplements Primitivism with the claim that our phenomenal states are not basic, but are instead the upshot of lower-level phenomenal goings-on that are beyond our immediate comprehension.

The panphenomenalist view does not qualify as a version of EV, but it might still be regarded as a promising account of the metaphysical status of consciousness. After all, we have already granted the inscrutability of the physical, so have opened the way for this position. However, there are at least two further considerations that count against this view.

First, we do not have positive reason to believe that inscrutables are phenomenal properties. We do have positive reason to believe that inscrutables are absolutely intrinsic properties, but not that they are subjective.[33] It could be argued that since our only concepts of absolutely intrinsic properties are *phenomenal* concepts, we should prefer the hypothesis that inscrutables are phenomenal over

[33] This is related to an objection to panphenomenalism that Seager (1995) labels the 'no-signs objection'. The idea is that basic physical entities do not display outward signs of being phenomenally conscious, so we should not attribute them phenomenal properties.

obscure claims of conceptual ignorance. This argument is unpersuasive. It is inappropriate to believe that inscrutables are phenomenal properties just because, as Coleman puts it, '…we can't be bothered to think any further than the ends of our minds.' (2008, p.89) This would reflect a deep lack of imagination and be highly anthropocentric. There is simply no need, independent of a desire to evade the –tivity gap, to regard inscrutables as subjectivity-involving. This speculation does not integrate with our existing understanding of the physical world.

Second, it is not clear that regarding inscrutables as phenomenal would successfully undermine the –tivity gap. Stoljar argues that '…it seems just as hard to see how one experiential truth can entail another as it is to see how a nonexperiential truth can entail an experiential truth' (2006, p.120). Block captures this in a thought-experiment. Imagine miniature conscious beings who '…build hordes of space ships of different varieties about the size of our electrons, protons, and other elementary particles, and fly the ships in such a way as to mimic the behaviour of these elementary particles' (1980, p.280). If our brains were constituted by such an alien horde, our brain would involve the instantiation of phenomenal properties, but *our* conscious experience would remain inexplicable. In other words, regarding inscrutables as inherently subjectivity-involving may not help account for the subjectivity of *our* conscious states.[34]

A last resort would be to take the 'neutral' route, and claim that inscrutables are neither objective nor subjective.[35] On this view, the principle that there is no entailment from the objective to the subjective becomes irrelevant. However, since inscrutables are not regarded as inherently subjectivity-involving, there is also no problematic commitment to them being phenomenal. Even if we permit the dubious notion of properties that are neither subjective nor objective, this position offers us the *worst* of both worlds. We have no reason to believe that neutral properties are any more suited to explaining subjectivity than objective properties are. Furthermore, if

[34] A more detailed objection to the idea of phenomenal 'atoms' explaining unified conscious states is offered by William James (1890). Chalmers describes this as '…easily the most serious objection…' to this kind of account (2002, p.267). We will consider James's argument further in Chapter 6 (Section 2.4).

[35] Such a position should not be confused with the 'neutral monism' of Mach (1886), James (1904), Russell (1910) and others.

neutral properties are to any degree '…tainted with the phenomenal…' (Montero 2010, p.77), all the objections against the panphenomenalist route arise once more.

CONCLUSION

Where does this leave RIH and EV? EV is defensible only if it offers a version of the ignorance hypothesis that satisfies the Relevance Condition and the Integration Condition. RIH goes a long way towards achieving this. The Inscrutability Argument makes a strong case for our being conceptually ignorant of the absolutely intrinsic properties of physical objects. Once we factor this ignorance into our reflections on consciousness, the –trinsicality gap loses its force. The hypothesis that inscrutables play a key role in the explanation of consciousness also satisfies the Integration Condition with ease. However, RIH has no force against the –tivity gap, so fails to satisfy half of the Relevance Condition. Consequently, RIH is not defensible.

Should the advocate of EV scrap RIH and pursue an *alternative* version of the ignorance hypothesis? We have learned two lessons from our examination of RIH that strongly suggest such a project would fail. One, it seems that EV can only evade the –trinsicality gap if the physical properties of which we are conceptually ignorant are *absolutely intrinsic* properties. RIH is the only version, or at least the only promising version, of the ignorance hypothesis that posits such properties. We thus have reason to doubt that any alternative version of the ignorance hypothesis could undermine the –trinsicality gap. Two, the considerations that cast doubt on RIH's ability to evade the –tivity gap will apply to *any* version of the ignorance hypothesis. Any proposed unconceived physical properties had better be *objective*, but it is doubtful that positing any kind of unknown objective property would undermine the principle that there is no entailment from the objective to the subjective.

Overall, our conclusion should not just be that RIH fails, but that EV is implausible. Though EV had a number of attractive characteristics, its success depends on whether it can deal with both the –trinsicality and –tivity gaps. We now have strong reason to believe that no version of the ignorance hypothesis will allow EV to achieve this.

CHAPTER 5

REPRESENTATIONALIST ACCOUNTS OF CONSCIOUSNESS

As of yet we have not found an adequate response to the Problem of Consciousness. EV showed great promise, but was ultimately found to be inadequate. The purpose of this chapter is to explore an alternative approach to overcoming the problem. Representationalists (or Intentionalists) assert that phenomenal states are intentional states. In having a conscious experience, necessarily we represent things to be some way or another. This relatively uncontroversial claim starts to have implications for the metaphysics of consciousness when two further claims are added. First, that the phenomenal properties of conscious states are *exhaustively determined* by their intentional properties (Strong Representationalism). Second, that the intentional properties of conscious states can be fully accounted for in *physical* terms (Physicalist Representationalism).[1] If these further claims can be defended, then the notion of representation would provide a bridge between the physical and the phenomenal, thus accommodating conscious experience in a Physicalist ontology.

Representationalism is a common position in the current debate over the Problem of Consciousness. In Chapter 2 I argued that standard responses to the problem are unsatisfactory, which motivated us to explore the non-standard response offered by EV. Why, then, are we now putting EV aside to reconsider an established position? I will ultimately conclude that the Representationalist strategy is indeed unsatisfactory. However, I will also argue that, like RIH, it succeeds in going *some* way towards addressing the epistemic gap, and so solving the Problem of Consciousness. RIH and the best form of Representationalism each deal with one half of the epistemic gap but not the other. This conclusion will allow me to develop a *hybrid account* in Chapter 6 that combines elements of RIH and of Representationalism to form a complex view of the phenomenal that overcomes the Problem of Consciousness. Nevertheless, in order to avoid prejudging the matter, the prospect of such a hybrid

[1] Physicalist Representationalism entails Strong Representationalism, but not vice versa.

will have no bearing on this chapter's assessment of Representationalism. I will evaluate Representationalism over three sections.

Section 1 will outline the Representationalist strategy and explain what is required for it to vindicate Physicalism. As will now be familiar, I distinguish two challenges: accounting for the *qualitative character* of a phenomenal state and accounting for the *subjectivity* of that state. In Section 2 I assess the prospects of Representationalism regarding the first challenge. I conclude that our old enemy, the –trinsicality gap, cannot be overcome by an appeal to intentional properties, so Representationalism cannot meet the first challenge. In Section 3 I consider whether the second challenge is more tractable. I argue that an appeal to the meta-intentional content of phenomenal states may indeed be able to account for the subjectivity of consciousness. A Self-Representationalist theory based on the work of Kriegel (2005/2009) addresses the –tivity gap. I conclude that Representationalism makes it plausible that *being a phenomenal state* is explicable in physical terms, but does not make it plausible that *the qualitative character* of phenomenal states is also explicable in physical terms.

SECTION 1

THE VARIETIES OF REPRESENTATIONALISM

In 1.1 I elucidate and support the view that all phenomenal states are representational. In 1.2 I distinguish between Weak and Strong Representationalism, and in 1.3 I distinguish between Physicalist and Nonphysicalist forms of Strong Representationalism. Only a Physicalist form of Strong Representationalism has the potential to offer an answer to the Problem of Consciousness. In 1.4 I outline how we should evaluate Representationalism's prospects of overcoming the anti-Physicalist arguments offered by Primitivists.

1.1. THE INTENTIONALITY OF CONSCIOUS STATES

What does it mean to say that phenomenal states are representations i.e. that they are intentional states? To be an intentional state is to be directed at the world – to be *of* or *about* something. Mental states are intentional states. Believing that you are sitting in a chair is directed at the state of affairs of your sitting in the chair. It represents the world as being a certain way. An intentional state and what it is about stand in an intentional relation. A special characteristic of intentional relations is that the represented thing need not obtain: there is a possibility of *mis*representation. For instance, one can believe that they are sitting in a chair when they are not. Non-intentional relations do not have this special feature. Sitting in a chair is a relation between you and the chair that can only hold if the chair exists.

What goes for mental states in general goes for *phenomenal* mental states in particular. Just as a state cannot be a belief unless it is a belief *about* something, a state cannot be a conscious experience unless it is an experience *of* something. It may not be of something *actual*, but it will always have intentional directedness. There is no such thing as an experience without intentional properties. As Tye claims, '*All* states that are phenomenally conscious - *all* feelings and experiences - have intentional content.' (1995, p.93). This claim is relatively uncontroversial, and several compelling arguments can be deployed in its favour.

First, the plausibility of the claim that *all* mental states are representations – that intentionality is the 'mark of the mental' – strongly suggests that all phenomenal states are representations (see Crane, 2007). Second, the terms with which we describe consciousness are representational terms. Phenomenal states are essentially states of *awareness*, and being aware is an intentional notion. Similarly, your experiential state is how things *seem* to you, how they *feel* or how they *appear*. Seeming, feeling and appearing are representational notions (see e.g. Byrne 2001, p.207). Furthermore, the transitive use of the word 'conscious' indicates that consciousness involves representation. Transitive consciousness is consciousness *of* – such as being conscious of your hunger – and consciousness *of* is a representational matter. Third, conscious states are the kind of thing that can be assessed for *accuracy*

– they can be veridical or non-veridical. It is only possible to assess something as being accurate or inaccurate if it has content i.e. if it is a representation. Overall, it is highly plausible that being an intentional state is a *necessary condition* of being a conscious state. Unsurprisingly, very few positions on consciousness deny this.

1.2. WEAK AND STRONG REPRESENTATIONALISM

We have concluded already that being an intentional state is a *necessary* condition of being a phenomenal state i.e. that all phenomenal states have intentional properties. Weak and Strong Representationalists both agree with this conclusion. Where they diverge is on the *sufficient* conditions of phenomenal property-instantiations. The difference between the two positions is as follows:

> ***Weak Representationalism:*** The phenomenal nature of a state is *not* exhaustively determined by the intentional properties of that state.
>
> ***Strong Representationalism:*** The phenomenal nature of a state *is* exhaustively determined by the intentional properties of that state.

In other words, only Strong Representationalists claim that whether a state is phenomenal, and what qualitative character that state has, is determined entirely by the intentional properties of that state. Only Weak Representationalists claim that non-intentional properties also contribute to whether a state is phenomenal and/or to the qualitative character of that state.

If Representationalism is going to overcome the Problem of Consciousness, it had better be *Strong* Representationalism. The promise of explaining intentional properties in physical terms only provides a vindication of Physicalism if conscious states can be *fully explained* in terms of intentional properties. As such, Weak Representationalism is not a strong enough thesis to provide what is needed. We will evaluate the prospects of Strong Representationalism over the course of the chapter. From here on, by 'Representationalism' I will mean specifically *Strong* Representationalism unless otherwise stated.

To appreciate the Representationalist claim that intentional properties determine phenomenal properties, we should distinguish between *pure* and *impure* intentional properties. Pure intentional properties are those that characterise the *content* of a representation. Impure intentional properties are those that characterise the *manner* in which that content is given. The content of a representation is what it says about the world – the conditions under which it is accurate. For example, your perceptual state might have the content that your leg is damaged. This perceptual state is accurate iff your leg is in fact damaged. According to some, two states with precisely the same content may nevertheless differ in another intentional respect; they might have the same content in a different *manner*. For instance, feeling that your leg is damaged and seeing that your leg is damaged might be mental states with the same content, but which have that content in different modes. Crane, for example, proposes that '…the difference between feeling one's leg to be damaged and seeing it to be damaged is just the difference between *feeling* and *seeing*…' (2007, p.12). The claim is that these two states have the same *pure* intentional properties, but differ with respect to their *impure* intentional properties.

This distinction between pure and impure intentional properties has ramifications for the evaluation of Representationalism. Say it transpires that the pure intentional properties of a state do not determine its phenomenal properties. In this case, one might still defend Representationalism by holding that a state's phenomenal properties are determined by its pure *and impure* intentional properties.

1.3. PHYSICALIST AND NONPHYSICALIST REPRESENTATIONALISM

If (Strong) Representationalism is true, then phenomenal properties are nothing over and above intentional properties. But are intentional properties nothing over and above physical properties? That is, can the intentional properties constitutive of a state's phenomenal nature be accounted for in physical terms? The two options are clear:

> ***Physicalist Representationalism:*** The intentional properties that determine a state's phenomenal nature *can* be accounted for in physical terms.
>
> ***Nonphysicalist Representationalism:*** The intentional properties that determine a state's phenomenal nature *cannot* be accounted for in physical terms.

The main reason for adopting Physicalist Representationalism is the thought that *all* intentional properties can be accounted for in physical terms. Here the philosophical debate surrounding consciousness coincides with that surrounding mental representation (see Hellie 2007, p.290). If we have reason to believe that all intentional properties of mental states are not onticly distinct from the physical, and we have reason to believe that conscious mental states are simply those with certain intentional properties, then we can conclude that consciousness is not onticly distinct from the physical. The reducibility of intentional properties in general motivates the reducibility of phenomenal intentional properties in particular.

Jackson, after his move away from Primitivism, captures the general outlook of Physicalist Representationalism:

> The project of finding an analysis of representation is not an easy one – to put it mildly. But...the answers that have been, or are likely to be, canvassed are all answers that would allow the fact of representation to follow a priori from the physical account of what our world is like. (quoted Davies 2008, p.27)

There is a widespread view that intentional properties can be accounted for in physical terms, but that precisely what those terms are has not yet been uncovered. As I have already argued (e.g. Chapter 3, Section 1.1.2), undermining the apparent inexplicability of consciousness in physical terms need not involve offering a positive physical explanation. It would suffice to argue that there *is* such an explanation even if we cannot specify *what* that explanation is. As such, it is acceptable to argue that consciousness can be explained in intentional terms amenable to Physicalism without offering a specific physical account of those intentional properties.

As an illustration of how intentional properties might be accounted for in physical terms, consider Tye's claim that 'S represents that P =df If optimal conditions obtain, S is tokened in x if and only if P and because P' (1995, p.101). Here S is the

vehicle of representation i.e. the state that represents, and P is the *content* of the representation i.e. what is represented by the state.[2] The formula describes the relation that must obtain between S and P in order for the former to represent the latter. Other theories offer more complex accounts of this representational relation, but the central idea remains the same: S and P, and the relation between them, need not involve the instantiation of any non-physical properties.[3]

Nonphysicalist Representationalists reject this reductive strategy. On this view, phenomenal states can be accounted for in intentional terms, but those intentional properties cannot be accounted for in physical terms (e.g. Chalmers, 2004). One route here is to claim that *no* intentional properties can be accounted for in physical terms. One might adopt this position if they believed that all intentionality is derivative of phenomenal intentionality (e.g. Searle, 1990). The irreducibility of phenomenal properties would then entail the irreducibility of all intentional properties. The more common route, however, is to accept that *most* intentional properties can be accounted for in physical terms, but claim that conscious states involve *special* intentional properties that are irreducible. The difference between non-phenomenal and phenomenal representations is that the latter involve onticly primitive phenomenal properties.

1.4. REPRESENTATIONALISM AND THE PROBLEM OF CONSCIOUSNESS

If Physicalist Representationalism can be defended, we would have a physical account of phenomenal consciousness. Since this position denies that there is an epistemic gap between the physical and the phenomenal, it qualifies as a *Type-A* position.[4] But how would such a position address the *anti*-Physicalist arguments for Primitivism that drive the Problem of Consciousness? When we first discussed Representationalism in Chapter 2, we noted the threat that the arguments against Physicalism can simply be

[2] On the vehicle/content distinction, see Dretske (2003 p.68) and Hutton (2009, p.21).
[3] Physical accounts of *impure* intentional properties have also been offered, often in functional terms.
[4] There are a number of Type-B versions of Representationalism (e.g. Tye 1995, p.180). However, such positions do not respect the A Priori Entailment Criterion established in Chapter 2, so will be disregarded.

re-applied to Physicalist Representationalism. As Alter puts it, '…bringing representationalism to bear on the debate over whether consciousness is physical leaves everything more or less as it was.' (2007, p.74) This worry applies to both CA and KA.

Regarding CA, Representationalism must claim that zombies and inverts are ultimately inconceivable. Once we realise that our physical duplicates have the same intentional properties as us, and that intentional properties fix phenomenal properties, it will be inconceivable that they should differ from us phenomenally. However, as Crane argues '…if the worry was that zombies seem to be possible, then (given the manifest possibility of non-conscious intentionality) this worry will arise even on [an] intentionalist approach.' (2007, p.24). The same goes for qualia inverts.

Regarding KA, Jackson (2007) has come to adopt Representationalism as a response to his own argument. The property of *representing* red in the manner required for a red-experience is, he claims, a property of which Mary has full knowledge in her monochromatic room. When she leaves the room, she comes to instantiate that property herself, but learns nothing new. However, here a critic can simply re-affirm the intuition that Mary *does* learn something new on leaving her room, indicating that the relevant representational property is *not* physical (see Alter, 2007).

How should we determine whether Physicalist Representationalism is refuted by the anti-Physicalist arguments, or overcomes them? We have established already that the plausibility of those arguments depends on the plausibility of the –trinsicality gap and the –tivity gap. The best way to assess the prospects of Physicalist Representationalism is to see how it fares with each of these gaps. Fortunately, the Representationalist literature generally distinguishes between accounting for the *qualitative character* and *subjective character* of consciousness, so facilitates this methodological division. We will evaluate whether a Representationalist account of qualitative character can address the –trinsicality gap, and whether a Representationalist account of subjective character can address the –tivity gap. If Representationalism succeeds in both tasks, then we can conclude that it overcomes

the epistemic gap at the heart of the Problem of Consciousness, and so undermines CA and KA.

SECTION 2

REPRESENTATIONALISM AND QUALITATIVE CHARACTER

In order to assess the prospects of a Physicalist Representationalist account of the qualitative character of conscious state, we must put aside the question of what makes a state one of subjective awareness. We will assume that a state meets whatever conditions (representational or otherwise) are required for it be a phenomenal state as such, and ask whether the intentional properties of that state plausibly determine *what it is like to be* in that state for its subject. The first step will be to consider whether a good case can be made for *Strong* Representationalism about qualitative character. I argue that it can. The second step will be to consider whether the intentional properties constitutive of qualitative character can be accounted for in terms amenable to *Physicalism*. I argue that no such account is available.

2.1. STRONG REPRESENTATIONALISM ABOUT QUALITATIVE CHARACTER

To take a Strong Representationalist view of qualitative character is to hold that *the intentional properties of a phenomenal state fully determine the qualitative character of that state*. What it is like for a subject to be in a conscious state is exhausted by the intentional properties of that state. What reasons are there to affirm this claim? Various arguments have been made, but the most compelling is the *argument from transparency*. A striking feature of the qualities of experience is their 'diaphanous' nature. Tye describes this feature: 'In turning one's mind inward to attend to…experience, one seems to end up concentrating on what is outside again, on external features or properties.' (1995, p.30) For instance, when we try to focus on the

qualitative redness of our experience as we perceive a postbox, we inevitably attend to the redness *of the postbox*. This plausibly applies to all phenomenal qualities. Whenever we introspect our phenomenal states, we only find the properties *represented* by that state rather than some property of the experiential state itself (Tye 1995, p.136).

The phenomenon of transparency invites a representational view of qualitative character. Our experience represents the postbox as red, and the properties that characterise our experience are *represented* properties rather than properties of the representation itself (see e.g. Dretske 2003, p.72). We see *through* our conscious states to their content. Even if we are imagining or hallucinating a red postbox, it remains the case that the redness is something our experience represents (albeit non-veridically), rather than something it instantiates. As Dretske proposes, '…experienced qualities, the way things phenomenally seem to be…are - all of them - properties the experience represents things as having.' (2003, p.67) In other words, the intentional properties of conscious states determine their qualitative character. Overall, the transparency of phenomenal qualities lends substantial support to Strong Representationalism about qualitative character.

Are there any plausible counter-examples to Strong Representationalism about qualitative character? A number of cases have been put forward in which pairs of experiences are claimed to have the same content, but to differ in their qualitative character.[5] For instance, *believing* that the cookie monster is blue and having a *mental image* of him as blue have the same content, but differ qualitatively since only the latter involves qualitative blueness (Kind 2007, p.406). Similarly, *seeing* something overhead and *hearing* something overhead are perceptual states with the same content, but what it is like to be in those states differs (Block 1995). Furthermore, one's visual experience before and after removing one's glasses could have the same content, but differ in qualitative character because the second experience is *blurry* (e.g. Crane 2007).

[5] There are also putative cases of experiences with qualitative character and no content, and experiences with the same qualitative character but which differ in their content. These examples are unpersuasive (see Kriegel 2009, Ch. 3).

One response to cases such as these is to claim that *impure* intentional properties contribute to qualitative character, and that the experience-pairs differ with respect to their impure properties. For instance, representing *blurrily* is a property of your experience that contributes to what it is like to be in that intentional state, as is representing *visually* (e.g. Crane 2007). Since such impure properties are still intentional properties, it remains the case that the intentional properties of a state fix its qualitative character. There are a number of serious objections to this 'adverbial' move (some of which we will consider in Section 2.2) but the most important objection is that it is *unnecessary*. The putative counter-examples can be dealt with by showing that there is in fact a difference in the *pure* intentional properties of the experience-pairs, so there is no need to confuse things with an appeal to impure properties.

The experience-pairs all have *overlapping* content, but it is not the case that they have the same *overall* content.[6] The mental image of the cookie monster represents him as having some fully specific shade of blue, where the belief that he is blue does not. This difference in content could plausibly account for the difference in qualitative character of the two representations. Seeing and hearing something overhead have some content in common - they each represent that an object is overhead. But the visual representation will inevitably contain very different information about that object than does the auditory representation. Finally, the 'glasses-on' experience represents things as having sharp edges while the 'glasses-off' experience represents them as having fuzzy edges (Dretske 2003, p.77). In the second experience, one might also be representing that the represented fuzziness is non-veridical; that the world is as the first experience represents it to be. But this would not mean that the overall content of both experiences is the same.

Though we have not dealt explicitly with all of the counter-examples raised against Strong Representationalism, we can tentatively conclude that it is accurate. Moreover, it is plausible that the *pure* intentional properties of a phenomenal state are sufficient to determine its qualitative character. The qualitative character of experience is simply how it represents things as being – how things seem to the subject of experience.

[6] This point is made with particular force by Dretske (2003, pp.75-80) and Kriegel (2009, pp.80-82)

2.2. PHYSICALIST REPRESENTATIONALISM ABOUT QUALITATIVE CHARACTER

Assuming that the qualitative character of conscious states is determined by their content, can the possession of such content plausibly be accounted for in exclusively physical terms? There are three components to a Physicalist account of qualitative content: the vehicle, the content, and the relation in virtue of which the vehicle has that content. Physicalism requires *all three* components to be explicable in physical terms. I will argue that though there are no specific problems pertaining to the vehicle component or relation component, there can be no physical account of the *content* of phenomenal representations. I will then deal with potential objections to this conclusion.

2.2.1. The Problem With Qualitative Content

Plausibly, there need be nothing special about the vehicle of a representation with qualitative content. If it is credible that physical states of the brain can act as vehicles of mental representation at all, then it should be credible that they can carry *qualitative* representations. Once we accept that phenomenal qualities are *represented* properties, the fact that brain states do not themselves have phenomenal qualities is not problematic. Something can represent a property without having that property itself. When we look into a subject's brain, we only see the vehicles of phenomenal representation. We should not expect knowledge of these vehicles to reveal the content of the representations they carry (see Dretske 2003, p.71). Conversely, the subject's experiential state will not reveal properties of the vehicle of that experience. They may be aware of qualitative redness, but it would be a mistake for them to infer that their brain instantiates red-qualities. Importantly, the –trinsicality gap presents no problem here. The fact that all properties of brain states are *structural* properties is compatible with their representing the *non-structural* properties that characterise our phenomenal states. Overall, there is no reason to think that the vehicles of qualitative content must be non-physical.

The relational properties that give a physical vehicle its content are a little more complex. We do not yet know what that relation involves. The causal covariance account outlined in Section 1, for instance, is clearly far too liberal. However, I have suggested it is plausible that the representation relation is ultimately explicable in physical terms. We cannot specify *what* relational properties are required to give brain states qualitative content, but they are presumably the same as the relational properties involved in an intentional state having *non*-qualitative content. The possession of qualitative content thus entails no *special* problem for Representationalism – only the same challenge of naturalising intentionality that applies to all mental representation. There are no compelling reasons to believe that the relation between a physical vehicle and its qualitative content is a non-physical relation.

This just leaves us with the *content* distinctive to qualitative representations. Presumably the Physicalist Representationalist must hold that the properties represented by our conscious states are physical properties of external objects. When our physical brain states stand in the right physical relation to these physical properties, they have qualitative content. The qualitative character of our experience is determined entirely by which physical properties the experience represents. In a veridical experience of the red post-box, the post-box bears the physical property of redness. In a non-veridical experience, it might be that no red-quality is instantiated, but it is nevertheless a physical property that the experience *mis*represents as occurring.

On this model, Representationalism is committed to *realism* about all of the qualities that characterise experience. If a brain state represents qualitative redness by standing in an appropriate relation to instances of qualitative redness, those instances had better be real mind-independent properties. Physical objects must literally be qualitatively red, in and of themselves. Of course, we might misrepresent the occurrence of such properties, but these properties must sometimes be instantiated in order for us to be able to represent them at all. Since Strong Representationalism must account for every aspect of qualitative character, this applies to *all* the qualities that contribute to our experience. Sounds, tactile feels, smells, tastes must all belong to

physical objects. Even pains must be physical features of our bodies rather than properties of experience. The list also extends to emotions, moods and perhaps intellectual experience.

For the sake of argument, assume that all phenomenal qualities are indeed real properties of external objects represented by conscious states. The problem for Physicalist Representationalism is that such qualities are not *physical* properties. We can motivate this conclusion by deploying the –trinsicality gap. Physical properties are exclusively structural, phenomenal qualities are non-structural, and there is no entailment from the structural to the non-structural. We originally formulated the –trinsicality gap on the assumption that phenomenal qualities are properties of *experience* rather than properties *represented* in experience. However, when Representationalists relocate those qualities outside the head, the –trinsicality gap loses none of its force. It applies to qualities-of-objects in just the same way as it applies to qualities-of-experience. For instance, a complete physical description of a post-box will not mention qualitative redness. Furthermore, the structural properties instantiated by the post-box cannot entail the instantiation of qualitative redness, as structural properties cannot entail non-structural properties. As such, the qualitative content of experience cannot be accounted for in purely physical terms.

2.2.2. Responses and Rebuttals

There are a number of possible responses to this objection to Physicalist Representationalism about qualitative character. I will outline some of these responses and show why each of them is unsatisfactory.

A) Qualities As Intrinsic Physical Properties: It could be maintained that the qualities of which we are aware in experience are actually physical properties. On this view, it is false that all physical properties are structural properties. The non-structural properties responsible for qualitative character are all *physical* non-structural properties. It may well be that *science* only describes the structural properties of physical objects, but consciousness reveals their non-structural features. As we saw in

Chapter 1, being mentioned in physical theory, whether current or ideal, is not a necessary condition of being a physical property (at least not on the sense of 'physical' that drives the Problem of Consciousness). Furthermore, as we saw in Chapter 4, being an extrinsic property is not plausibly a necessary condition of being a physical property. Of course, being non-phenomenal *is* a necessary condition of being physical, but the view in question does not posit qualities that are themselves phenomenal. After all, to attribute qualitative redness to a post-box is not to attribute it phenomenal consciousness.

The main problem with this view is that there is no plausible account of the causal status of such properties. Qualities simply do not figure in our causal explanations of physical events. It could be argued that such explanations are mistaken, or at least seriously incomplete, but this would be implausible. The redness of the post-box does not *do* anything; the post-box could be qualitatively green and have precisely the same causal profile. Physical events can be explained in purely structural terms.[7] One option is to concede that qualities are inefficacious. On this view, science describes the causal structure of the physical world, but conscious experience reveals some of the world's *inefficacious* physical properties. The problem here is that it is implausible that we could come to represent inefficacious properties. We may not have a complete theory of the representation relation, but causation is clearly going to play a major role. How could we come to represent qualitative redness if that property never had any affect upon us?

B) Qualities as Structural Physical Properties: An alternative strategy is to concede that all physical properties are structural (with the possible exception of inscrutables) but hold that properties represented by phenomenal states are actually *structural* physical properties. For example, Tye (2000) suggests that red experiences represent objects as belonging to a certain 'spectral reflectance class' and Hill (2004) argues that pain

[7] A qualification is needed here. In Chapter 4 I argued that there is a sense in which scientific explanations are incomplete; they describe the causal structure of events but not the absolutely intrinsic properties that ground that structure. However, inscrutables can be integrated into our existing understanding of causation in a way that sensory qualities cannot. For instance, attributing qualitative redness to a post-box would not fill a hole in our metaphysical understanding of its causal properties.

experiences represent bodily disturbances. In both cases, the represented property is a robustly physical structural property. On this view, the properties responsible for qualitative character are all properties that can be accommodated within physical theory.

Spectral reflectances are indeed physical properties, but they are not the kind of property representation of which could constitute a red experience. Qualitative redness simply has nothing to do with that physical property. If red experiences *do* represent this physical property, we are left with the mystery of why states that represent that property have the qualitative character they have. In other words, it becomes implausible that the content of a phenomenal representation determines its qualitative character. The same goes for the bodily disturbance view of pain – the qualitative feel of pain is simply ignored. There is an inevitable gap between an experience representing some structural physical property and its having some qualitative character.[8]

It could be held that we are *mistaken* that phenomenal qualities are intrinsic properties. After all, there are cases of properties that *seemed* to be intrinsic properties but transpired not to be. For instance, the property 'weight' seemed intrinsic but turned out to be relational. Maybe it will transpire that qualitative redness is a structural property, which would remove the apparent obstacle to its being a physical property of external objects.[9] It is implausible, however, that phenomenal qualities can be analysed in structural terms. When apparently intrinsic properties turn out not to be intrinsic, their analysis usually leaves an intrinsic 'residue'. In the case of 'weight', we are left with the intrinsic property of 'mass'. The prospects of analysing an apparently (absolutely) intrinsic property into *purely* structural terms are poor (see Levine 2001, p.101).

A possible response here is to make an appeal to *impure* intentional properties. On this view, part of what determines qualitative character is the *manner* in which our

[8] Such a gap may be acceptable for Type-B theorists, but we are pursuing a Type-A version of Representationalism.

[9] For a more sophisticated version of this claim of misrepresentation, see Pereboom (2011). His account faces objections similar to those discussed presently and which I consider more closely in McClelland (forthcoming).

phenomenal states represent. The thought is that phenomenal states represent robustly physical properties, such as spectral reflectances and bodily disturbances, but their qualitative character is fixed by the *way* in which those properties are represented. For instance, red experiences represent the relevant spectral property 'redly'. It is this adverbial property that gives a red experience its distinctive qualitative feel. This deeply implausible position fails for a number of reasons. One, it ignores transparency. Redness appears to us as a property of the object we represent, but this view makes it a property of the representation. If we do not take transparency seriously, the whole case for Representationalism falls apart. Two, when we perceive a red post-box next to a black bin, the adverbial account says we are representing 'redly' and 'blackly'. But no account can be provided of the fact that it is the *post-box* that appears red and the *bin* that appears black (see Crane, 2005). Third, we are left with the mystery of how representing *redly* can be explained in physical terms. Redness again becomes a property of experience, and we are faced with the problem that this property cannot be accounted for in terms of the exclusively structural physical properties of brain states.

C) Qualities as Non-Actual: A final radical option is to claim that phenomenal states represent non-physical properties, but that those properties are not instantiated in the real world. Our red experience represents qualitative redness, and that intrinsic property is non-physical. But since that property is *non*-actual, this is consistent with all *actual* properties being physical. Qualitative content is never veridical – all phenomenal states misattribute qualities to physical objects. The property of *representing* some phenomenal quality can be accounted for in physical terms, and this is enough to accommodate qualitative character into a Physicalist ontology.

The main problem with this line of thought is that there is no plausible account of how we could come to represent properties that are never instantiated.[10] Non-actual properties are inefficacious, so if the representation relation has a causal component, we cannot represent them. Of course, sometimes we represent things that are not really there. But it is one thing to sometimes misrepresent the occurrence

[10] See Shoemaker, quoted Pereboom (2011, pp.41-42).

of a property, and quite another to *never* represent it accurately. Remember, Physicalist Representationalism requires that the representation relation can be accounted for in physical terms. There is plausibly a physical account of how we represent properties that are sometimes actually instantiated, but it is implausible that there could be a physical account of how we represent non-physical properties that are never instantiated.

Overall, Physicalist Representationalism about qualitative character is implausible. The major threat to Physicalist Representationalism is that its reconceptualisation of consciousness leaves the anti-Physicalist arguments exactly as they are. We can now see how a Representationalist view of qualitative character asks us to reconceive phenomenal qualities as represented properties, but still falls victim to the –trinsicality gap. The intrinsic properties that characterise our phenomenal states cannot be accounted for in structural physical terms, even if those intrinsic properties belong to represented objects rather than to the phenomenal state itself.

SECTION 3

REPRESENTATIONALISM AND SUBJECTIVE CHARACTER

A state has subjectivity iff there is something it's like to be in that state for its subject. Is there a representational account of subjectivity amenable to Physicalism? To answer this question, we must put the issue of qualitative character aside. The question here is whether a Physicalist Representationalist account can be given of what makes something a conscious state at all, not of what makes it the kind of conscious state it is. All my mental states occur *in me*, but only my conscious mental states occur *for me* (Levine 2001, pp.6-7). They have a distinctive *presentedness* that must be accounted for. In the previous section I tentatively concluded that the qualitative character of phenomenal states is determined by their intentional properties, though those intentional properties cannot be accounted for in standard physical terms. Presumably

the intentional properties responsible for qualitative character can also be instantiated by non-conscious mental states. Our task is thus to discern whether the difference between *conscious* qualitative representations and *non*-conscious qualitative representations is a *representational* difference.

Interestingly, many Representationalist theories attempt to account for qualitative character in intentional terms, but *do not* attempt to account for subjectivity in intentional terms. Tye (1995), for instance, claims that a key difference between conscious and non-conscious states with qualitative content, is that conscious states are *poised* for cognitive processing. Similarly, Jackson (2007) cites the special role that conscious states play in belief-formation. There are two things to note about accounts of this kind. One, they are not *representational* accounts of subjectivity. They claim that phenomenal states are representations and that their qualitative character is determined by their content, but the feature that makes them phenomenal *is not* an intentional property. Tye and Jackson, for instance, are citing the *functional* status of a qualitative representation to account for its subjectivity. Two, this kind of account has dim prospects of defending Physicalism about subjectivity (see Kriegel 2002, pp.62-63 and 2009, p.72). We have already concluded in Chapter 2 that functionalist accounts of consciousness fail: the fact that the positions under discussion only try to explain the *subjective* aspect of phenomenal states in functional terms does not make them any more plausible. We will still, for instance, be able to conceive of a being whose qualitative representation performs the proposed functional role, but who lacks subjective awareness. This is another case in which anti-Physicalist arguments can simply be re-applied to a (so-called) Representationalist theory.

We are not concerned with the modest claim that subjective states are representational states. We are concerned with the bolder claim that it is the intentional properties of a state that *make it* a subjective state at all i.e. Strong Representationalism about subjectivity. If those intentional properties can plausibly be accounted for in physical terms, we would then have a Physicalist Representationalist view of subjectivity. Of course, the litmus test for such a position is whether it can successfully undermine the apparent conceptual gap between the objective and the subjective. I will consider the prospects for Physicalist Representationalism about

subjective character in three stages. In 3.1 I evaluate Higher-Order Representation (HOR) theory, arguing that it has significant promise but ultimately fails. In 3.2 I advocate Kriegel's Self-Representationalist account, which retains the virtues of HOR theory but avoids its weaknesses. In 3.3, I argue that this kind of Self-Representationalist account of subjectivity is plausibly amenable to a Physicalist ontology.

3.1. HIGHER-ORDER REPRESENTATION (HOR) THEORY

3.1.1. The Case for HOR Theory

The HOR theorist claims that a state is conscious iff it is the object of a distinct higher-order representation that one is in that state.[11] For instance, the perceptual representation of a post-box becomes a conscious state when it is appropriately represented by some further mental state – a representation with that *first-order* representation (FOR) as its object. A terminological clarification is needed: a subject is *transitively* conscious *of* their FOR in virtue of their HOR and, on the other side of the conceptual coin, their FOR is thereby an *intransitively* conscious state. When I talk of 'the conscious state', I will be referring to the intransitively conscious first-order state.

The motivation behind HOR theory is compelling. The difference between conscious and non-conscious states is that we are *aware of* our conscious states, and unaware of the non-conscious representations that occur within us (see e.g. Lycan 2004, p.93). Since awareness is a representational notion, to be aware of a mental state is to *represent* that mental state in some way. This encourages a straight-forward argument for HOR theory, usefully captured by Lycan:

> (1) A conscious state is a mental state whose subject is aware of being in it. [Definition]
>
> (2) The 'of' in (1) is the 'of' of intentionality; what one is aware of is an intentional object of the awareness.

[11] Often HOR theories are presented as 'theories of consciousness' rather than theories of the *subjective aspect* of phenomenal states. However, it is clear that those theories are best read as accounts of subjective awareness specifically (see Kriegel 2009, pp.114-115).

> (3) Intentionality is representation; a state has a thing as its intentional object only if it represents that thing.
>
> Therefore,
>
> (4) Awareness of a mental state is a representation of that state. [2,3]
>
> And therefore,
>
> (5) A conscious state is a state that is itself represented by another of the subject's mental states. [1,4] (2001, pp.3-4)

This argument is quite persuasive, and offers a clear account of the difference between conscious and non-conscious mental states. Subjective awareness involves intentional properties, though it does not involve intentional properties *internal* to the mental state of which we are aware. Rather, it is the HOR's intentional property of representing the first-order mental state that is responsible for our subjective awareness. When a mental state M makes the transition from non-conscious to conscious, its internal properties remain the same, but it acquires the *relational* property of being represented by some distinct mental state M*.

The argument above indicates that being the object of an HOR is a *necessary* condition of being conscious, but does not show that it is a *sufficient* condition. Indeed, all versions of HOR theory offer extra conditions that must be met for a state to be conscious. Chief among these are a condition of (rough) *simultaneity* between M and M*; M cannot be made conscious by you representing it after its occurrence, and of *non-inferentiality*; if you infer the presence of M, say through psychotherapy, that does not make M a conscious state.[12] Nevertheless, it is the condition that M is represented by M* that is meant to be doing most of the explanatory work. The task of supplementing that condition with *further* conditions that rule-out potential counter-examples is not of primary philosophical importance.

There are two prominent controversies that can be put aside for current purposes. First, there is some disagreement between HOR theorists about the nature of the HOR states responsible for consciousness. Higher-Order *Thought* theorists, such as Rosenthal (1986/2004), claim that M* is thought-like where Higher-Order *Perception* theorists, such as Lycan (1996/2004), claim that M* is more akin to a

[12] These further properties could be regarded as non-intentional properties or as *impure* intentional properties. Little seems to hang on this decision, though Van Gulick (2004) argues that these properties are *non*-intentional, and that their being so goes against the spirit of HOR theory.

perceptual representation. Closer scrutiny suggests that the difference between these two camps may not be as substantive as the debate would suggest, so there is no need to adjudicate between them (see Gennaro 1996).

Second, there is also disagreement about how qualitative character fits into the story. On one view, the qualitative character of a conscious state is determined by the intentional properties of M. On the alternative view, it is fixed by the intentional properties that M* *represents* M to have. Levine (2001, p.108) and others have argued that neither route is defensible.[13] However, this need not concern us. We want to know whether HOR theory can provide a plausible account of the *subjective* character of phenomenal states. Whether and how it can then be integrated with an account of qualitative character is important in the limit, but should not distract us here.

3.1.2. The Case Against HOR Theory

Despite HOR theory's initial appeal, it faces a number of objections. I will not address the various objections that pertain to how qualitative character fits into the account since, as I explained above, this is not our primary concern in evaluating HOR theory. Nevertheless, there are at least three serious objections to HOR theory as an account of subjective character: the targetless HOR problem, the generality problem and the self-intimation problem. In light of these, I conclude that HOR theory is implausible.

A) The Targetless HOR Problem: Sometimes representations *mis*represent.[14] More specifically, sometimes they represent something to exist when, in reality, there is no such thing. What goes for representations in general also goes for higher-order mental representations. This means it should be possible for a subject S to have a second-order state M* that represents the occurrence of M, even though S is not actually in

[13] The issue is also discussed informatively by Neander (1998).
[14] Indeed, the possibility of *mis*representation is often taken to be a necessary condition of something's being a representation at all (see Dretske, 1986).

M. Take it that S has no higher-order representations besides M*. The question is this: is S in a conscious state or not?[15] This question generates a dilemma for HOR theory.

On the one hand, it could be claimed that S is conscious. On this view, possessing an HOR is sufficient for being in a state of subjective awareness. But then which of S's mental states is a conscious state? Since S is not actually in M, it cannot be M that has the property of being a conscious state. Since M* is not the object of an HOR then, according to the theory, M* cannot be a conscious state either. To say that some *other* mental state of S is conscious would be wildly *ad hoc*, and could be ruled out by stipulating that S has *no* mental states besides M*. Overall, it is implausible that targetless HORs are sufficient for subjective awareness.

On the other hand, it could be claimed that S is *not* conscious. On this view, possessing an HOR is not sufficient for being in a state of subjective awareness: the HOR must also have a *target*. This approach holds, as it should, that S is not conscious, but does so at the expense of a central tenet of the theory. If HORs are meant to be responsible for subjective awareness, then S's possession of the HOR should be sufficient for her being in a phenomenal state. HOR theory does not attribute first-order states any role in explaining subjectivity, so why should M's absence prevent S from being in a state of subjective awareness? Overall, if targetless HORs are not sufficient for subjective awareness, HOR theory reneges on its promise of explaining subjectivity in terms of higher-order representation. This dilemma casts serious doubt on HOR theory. A number of responses to this objection have been put forward, but none are persuasive.[16]

B) The Generality Problem:[17] According to HOR theory, M is conscious in virtue of its relational property of being represented by M*. The problem here is that the relational property of being represented by a mental state is possessed by any number of non-

[15] Standardly, this objection is phrased in terms of whether *what* it is like to be S is the same as *what* it is like to be a subject in the same higher-order state as S, but who is also in M (e.g. Kriegel 2009, pp.129-139). This formulation muddies the waters as it brings qualitative character into the discussion. My version focuses on subjective character. Furthermore, there is another objection to HOR theory involving M* misrepresenting the qualitative character of the (actual) mental state M (e.g. Levine, 2001). Again, this pertains to qualitative character rather than subjectivity, so will not concern us.

[16] Kriegel (2009, pp.132-135) reviews and criticises a number of such responses.

[17] Proponents of this objection include Goldman (1993), Van Gulick (2004), Gennaro (2004) and Kriegel (2009, pp.143-4).

mental states. Presumably the HOR theorist will not hold that a rock becomes conscious when represented by a mental state (see Goldman, 1993). But if that relational property does not make the *rock* conscious, why should it make the *mental state* conscious? Claiming that only mental states can be conscious won't do. As I noted in the previous objection, according to HOR theory the represented state plays no role in the explanation of subjectivity. As such, its being non-mental should not make a difference to whether or not it is conscious. We need a non-*ad hoc* reason to limit consciousness to mental states, but HOR theory provides none.

C) The Self-Intimation Problem: The final objection to HOR theory I will consider is that it fails to do justice to an aspect of our phenomenology. The innocuous claim that drives HOR theory is that for all conscious states, we are aware of those states. The problem at hand concerns the further claim that for all conscious states, we are also *aware of our awareness*. To have a subjective state involves being conscious of our consciousness. On this view, subjectivity is not just the property *responsible for* subjective awareness; it also *falls within the scope* of that awareness. Call this the Self-Intimation thesis. It is hard to offer philosophical arguments for a phenomenological claim, but this thesis should be plausible on reflection.[18]

The Self-Intimation thesis is widely accepted within the phenomenological tradition, and is referred to as *pre-reflective self-awareness*.[19] The thesis also has a number of adherents in the analytic tradition. Strawson holds that '…all awareness comports awareness of itself…' (2011, p.282) and goes on to quote an illuminating passage from Frankfurt: '…what would it be like to be conscious of something without being aware of this consciousness? It would mean having an experience with no awareness whatever of its occurrence. This would be, precisely, a case of unconscious experience.' (quoted Strawson 2011, p.285) Note, the claim is not that all consciousness requires *introspection*. Nor that consciousness requires a rich 'egological' awareness of ourselves as persons (see Kriegel 2009, p.178). Nor that this

[18] Kriegel (2009, pp.113-129) makes the interesting argument that HOR theory presupposes that awareness is always phenomenally manifest, since this is the only plausible source of evidence for its central claim that all conscious states are represented states. However, Kriegel's case against *non*-phenomenal evidence that conscious states are always represented is unpersuasive (see Van Gulick, 2011).

[19] For a summary see Kriegel (2009, p.176).

self-awareness is a focal feature of our experience (see Kriegel 2009, p.190). Rather, the claim is just that the awareness constitutive of M's being conscious is itself phenomenally manifest to the subject.

Assuming that the Self-Intimation thesis is true, why should it present a problem for HOR theory? According to HOR theory, to be aware of M requires a distinct state M* in virtue of which we are aware. To be aware *of our awareness* of M, we would need some third state M** that represents M*. Such higher-order representation would then make M* a conscious state. But, according to the Self-Intimation thesis, in all conscious states we are aware of our awareness. For our awareness of M* to be phenomenally manifest, we would need a *fourth* state. Now an infinite regress looms, or perhaps a vicious circle.[20] Overall, HOR theory cannot account for the self-intimation of subjective states.

3.2. SELF-REPRESENTATIONALISM

3.2.1. Self-Representationalism About Subjectivity

Various positions are available that retain the central insight of HOR theory whilst avoiding the three objections above. HOR theory is right to claim that to be *aware* of a mental state is to *represent* that state in some way. However, we can challenge its presupposition that the representing and represented states are *distinct* states. Looking back at Lycan's argument for HOR theory in Section 3.1.1, the conclusion that the FOR and HOR states are distinct does not follow from the premises. Van Gulick (2006) labels this *the distinctness assumption*, and it is plausibly this assumption that makes HOR theory vulnerable to the three objections discussed. 'One-state' theories reject the distinctness assumption. They claim that conscious states have first-order content *and* have the meta-intentional content required for subjective awareness. No distinct higher-order representation is involved.

[20] See Kriegel (2009, pp.124-125).

One-state theories are offered by Gennaro (2004), Van Gulick (2004, 2006) and Kriegel (2005, 2009). We will focus on Kriegel's position as the strongest and most thoroughly-developed, though there may well be some respects in which the competing positions have advantages. Kriegel's core claim is that '…what makes something a conscious state at all, what constitutes its subjective character, is a certain kind of *self-representation*…' (2009, p.13). Call this 'Self-Representationalism'.[21] On this view, a mental state has subjectivity in virtue of suitably representing itself. How does an account along these lines avoid the three objections to HOR theory?

A) The Targetless HOR Problem: This problem rests on the possibility of an HOR representing a mental state to occur when no such mental state exists. This kind of misrepresentation is impossible on the self-representational account. If M is a self-representational state, then it is an *actual* state. M cannot represent itself to exist *and be wrong*. It might misrepresent itself in *other* ways, but that's beside the point.[22] By denying the distinctness assumption, Self-Representationalism excludes the possibility of targetless HORs. When meta-intentional content is *reflexive*, it cannot fail to have a target.

B) The Generality Problem: This problem rests on HOR theory's commitment to consciousness being a *relational* property that, it seems, is easily possessed by non-mental states. According to Self-Representationalism, being a conscious state does *not* consist in standing in some relation to a distinct state. Rather, it is a property *internal* to conscious states. Being *self*-representing is much more demanding than simply being represented. Rocks and other non-mental entities simply cannot self-represent (or at least cannot do so in the right way). As such, Self-Representationalism avoids any commitment to non-mental states being states of subjective awareness (Kriegel 2009, p.145).

---

[21] Kriegel reserves the label 'Self-Representationalism' for a position that combines this account of subjective character with his distinctive account of *qualitative* character (2009, p.165). My use of the term does not involve this second commitment.

[22] M might represent itself to have one qualitative character when in fact it has another. This will not concern us, though for further discussion see Kriegel (2009, pp.137-8) and, for an alternative position, Van Gulick (2006).

C) The Self-Intimation Problem: When HOR theory attempts to account for the self-intimation of conscious states, it generates an infinite regress (or vicious circle). Self-Representationalism halts that regress before it starts. All awareness comports awareness of itself because all states of awareness are self-representing states. That is, the awareness is internal to the conscious state itself. On this view, no extra state is required to account for our self-awareness, so no regress looms. Self-Representationalism respects the thesis that all conscious states are self-intimating (Kriegel 2009, pp.197).

In summation, Self-Representationalism overcomes the three objections raised against HOR theory. Moreover, it does so in a particularly *straightforward* way. This is important dialectically: the challenge for the HOR theorist is not just to respond to the three objections, but to do so in a way that is preferable to the clear and simple solution offered by Self-Representationalism.

Self-representation is plausibly a necessary condition of being a conscious state. Can we build up to a sufficient condition? Kriegel claims that a state is conscious if it *suitably* represents itself, and offers three criteria that flesh out this suitability clause. These criteria rule out a number of potential counter-examples to Self-Representationalism.

First, a subjective state must be *non-derivatively* self-representing (2009, p.158). That is, the state must have its content independently of interpretation. The sentence 'this very sentence is in English' is self-representational, but its meaning is derivative from interpretation, so does not entail consciousness. Second, a subjective state must be *specifically* self-representing (2009, p.159). An unconscious belief that all beliefs are neurophysiologically realised is self-representing in the sense that it is part of its own extension. Subjective awareness requires that the self-representation purports to represent a specific particular. Third, a subjective state must be *essentially* self-representing (2009, p.161). Rather than just happening to have itself as a referent, it must represent itself *as* itself. The contrast here is analogous to Perry's (1979) famous contrast between thinking 'the person with the torn bag is making a mess' and thinking '*I* am making a mess'.

With these three conditions, a serious proposal can be made about the necessary and *sufficient* conditions of subjective awareness. Self-Representationalism about the subjective character of phenomenal states is the following thesis:

> Necessarily, for any mental state M, M has subjective character iff M is non-derivatively, specifically, and essentially self-representing. (Kriegel 2009, p.164)

The central thought behind this account is that subjectivity is awareness, that awareness requires representation, and that this representation had better be reflexive representation. There are also a number of further motivations for Self-Representationalism that I will not discuss. Overall, Self-Representationalism has significant promise, and is the best candidate available for a representational account of subjectivity.

3.2.2. Self-Representationalism and the Anti-Physicalist Arguments

Physicalist Representationalism about subjective character requires subjectivity to be explicable in intentional terms, and those intentional properties to be explicable in physical terms. The second step is unlikely to generate any special problems for Self-Representationalism. If it is accepted that mental representation in general is explicable in physical terms, it should be accepted that non-derivative, specific, and essential *self*-representation is explicable in physical terms.[23] The challenge for Self-Representationalism is to defend the claim that subjectivity requires nothing more than the instantiation of those physically-realisable intentional properties.

Levine (2006) maintains that there is an epistemic gap between being in a suitable self-representing state and being in a state of subjective awareness. This gap is evident in the fact that we can conceive of a being who has the proposed self-representing mental states, but who is nevertheless not conscious (see Van Gulick 2011). This simply re-applies the zombie version of CA to Self-Representationalism.

[23] One potential objection to this claim is that physical accounts of representation cannot plausibly accommodate *reflexive* representation. This worry is dealt with by Kriegel (2009, Ch. 6), and it would take us too far astray to discuss it here.

None of this is to say that self-representation is not *necessary* to subjectivity, but it does cast doubt on the *sufficiency* claim needed to vindicate Physicalism.

One option for the Self-Representationalist is to suggest that the epistemic gap would disappear under ideal epistemic circumstances and, accordingly, that zombies are not ideally conceivable. After all, we do not yet have the right physical theory of representation. Perhaps our intuitions about the epistemic gap will change once we do have the right theory and see precisely how physical properties can carry subjective representations.[24] Perhaps we cannot really conceive of zombies with the relevant self-representing states. Without a theory of representation, how do we even know what to imagine?

Of course, anti-Physicalists can simply respond that any future discovery will offer the *wrong kind* of information to account for subjectivity. This is where the –tivity gap comes in: all physical properties are objective, and there is no entailment from the objective to the subjective. The intentional properties carried by physical states are just more objective properties - properties the instantiation of which does not entail consciousness. Perhaps there is a *special* kind of representation that does suffice for subjective awareness, but this is not the kind of representation that could ever fully be accounted for in terms of objective properties.

Here we see that the crux of the anti-Physicalist position is the –tivity gap. If Self-Representationalism can cast doubt on that apparent conceptual gap, Physicalism can be protected. Remember, the task is to undermine the claim that subjective awareness is *inexplicable* in physical terms, not to actually offer a complete physical explanation of subjectivity. The core of Self-Representationalism's challenge to the –tivity gap is that once we understand subjectivity representationally, the objectivity of physical properties no longer rules out the possibility of physical explanation.

It is highly plausible that subjective states are representational states. This much is clear from the simple fact that consciousness involves awareness, and awareness is representational. All representations have vehicles, and we know that vehicles can be quite unlike the representations that they carry. Just as a vehicle need

[24] Note, since this response does not involve conceptual ignorance, it is a version of the 'rudimentary response' to Primitivism discussed in Chapter 1 (Section 3.1) rather than a version of EV.

not be red to carry the intentional property of representing redness, it need not be inherently subjectivity-involving to carry the intentional properties responsible for subjectivity. Does the –tivity gap really cast any doubt on this? It is implausible that undergoing a subjective representation gives us some deep insight into what kind of vehicle that representation must have. It may *seem* to us that the state we are in is not constituted by objective properties, but this is just a run-of-the-mill case of confusing intentional properties for vehicular properties. When we reflect on the redness of our experience, it is sometimes hard to believe that the state we are in is fully implemented by properties that do not involve qualitative redness. Similarly, when we reflect on the subjectivity of experience, it is sometimes hard to believe that the state we are in is fully implemented by properties that do not themselves involve subjective awareness. We should not trust our intuitions in either case.

There are no compelling reasons to believe that subjective states can only be implemented by properties that are not objective. It is plausible that the physical vehicles of self-representation are robustly objective and carry the intentional properties constitutive of subjective awareness without the addition of any onticly primitive subjective ingredient. Given an appropriate theory of representation, there would be an a priori entailment from the relevant objective properties of the vehicle to the subjectivity of the representation it carries.

This casts significant doubt on the –tivity gap, but is unlikely to persuade the anti-Physicalist. The intuition that there is no entailment from the objective to the subjective is bound to persist. To reinforce my case against the –tivity gap, I will try to *explain away* that intuition: to show why the –tivity gap seems plausible despite being false. To do this, I adapt an argument offered by Kriegel (2009, pp.289-298).

Self-Representationalism claims that objective properties can constitute subjective awareness by performing a particular role. Specifically, the vehicular role of carrying the right kind of mental representation. To advocate the –tivity gap is to deny that there is any role the performance of which is sufficient for the instantiation of subjectivity. Water is reducible to H_2O because there is a role – 'the water role' – performance of which is sufficient for being water, and which is in fact performed by H_2O (as discussed Chapter 2, Section 2.2.1). By contrast, subjectivity appears

irreducible to objective properties because no matter what role is performed by an objective state, it remains an open question whether it constitutes a state of subjective awareness.

This appearance is at the heart of the Problem of Consciousness, and was explored in Chapter 1. The key questions is this: *why* does subjectivity appear to be something over and above the performance of some role? A natural suggestion is that it appears so because *it is* so. However, an alternative suggestion is that it only appears so because of the distinctive epistemic status of subjectivity. On this view, even if the performance of some role *is* sufficient for subjectivity, it would still appear to us that it is *not*.

Due to the Receptivity of knowledge, as discussed in Chapter 4, we generally know properties via their causal role. If our epistemic access to a property F is limited to its causal manifestations, then our criteria for being F are inevitably limited to the performance of some causal role. This makes F appear open for reduction to whatever property performs that role. Unlike other properties, subjectivity is not something we know via its causal role. Because conscious states are self-representing, our epistemic access to consciousness is built into the very instantiation of consciousness. As Kriegel puts it, '…knowledge of consciousness, and of consciousness alone, does not require causal contact with the known.' (2009, p.295)

Now, given that we do not *access* subjectivity via its causal role, we are liable to think that subjectivity consists in something *more than* the performance of some role. Subjectivity is the one property that we know *supra-causally*, which makes it appear to be a supra-causal property. But things would seem this way even if subjectivity did in fact consist in the performance of some appropriate role – a role performed by objective properties. As such, the appearance cannot be taken at face value. It does not follow from the fact that our *criteria* of subjectivity are role-transcendent, that subjectivity itself is metaphysically role-transcendent.

Self-Representationalism thus offers an explanation of why subjectivity *appears* inexplicable in objective terms that is compatible with it in fact being explicable in objective terms. This alone does not show that subjectivity *is* nothing over and above certain objective properties, but it removes the main reason to doubt that this is so. It

is defensible to reject the apparent –tivity gap and claim that subjectivity requires nothing more than the performance of the right role by objective properties; specifically, the role of carrying a suitable self-representing mental state.

CONCLUSION

Where has this brief foray into Representationalism left us? Representationalism offers a solution to the Problem of Consciousness only if phenomenal consciousness can be accounted for in terms of intentional properties, and those intentional properties can be accounted for in terms of physical properties. The qualitative character of phenomenal states might be intentionally determined, but attempts to account for the relevant intentional properties are vulnerable to the –trinsicality gap. The subjective character of phenomenal states is plausibly the upshot of those states suitably representing themselves. Moreover, this Self-Representationalist account casts serious doubt on the –tivity gap. Overall, Representationalism can only go *half way* to undermining the epistemic gap that drives the case for Primitivism. Consequently it fails to offer a comprehensive response to the Problem of Consciousness.

CHAPTER 6

THE NEO-RUSSELLIAN IGNORANCE HYPOTHESIS

In Chapter 1 we saw that at the heart of the Problem of Consciousness is the apparent epistemic gap between the physical and the phenomenal, which breaks down into the –trinsicality and –tivity gaps. In Chapters 3 and 4 we explored the Epistemic View (EV). We saw that the Russellian Ignorance Hypothesis (RIH), according to which we have no conception of the inscrutable intrinsic nature of physical entities, undermined the –trinsicality gap but fell victim to the –tivity gap. In Chapter 5 we considered a Representationalist approach to the problem. We found that no version of Representationalism could avoid the –trinsicality gap, but suggested that a plausible Self-Representationalist account of subjectivity could confront the –tivity gap. From this, our next line of enquiry should be clear: is there a way of *combining* RIH and Self-Representationalism to form a *hybrid* account of phenomenal consciousness? If so, we should be able to address *both* of the two apparent conceptual gaps, and so overcome the Problem of Consciousness.

In this chapter I will present and defend just such a hybrid position, which I call the Neo-Russellian Ignorance Hypothesis (NRIH). In Section 1 I will present the fundamentals of NRIH, and consider the dialectical situation in which we find ourselves. In Section 2, which constitutes the bulk of the chapter, I consider a range of potential objections. NRIH entails a number of commitments about the explanatory basis of consciousness. The five problems I consider all threaten to show that there remains an *epistemic gap* between the proposed explanatory base and consciousness. I argue that these problems can be overcome and, in the process, fill in some details of NRIH. In Section 3 I explain precisely how NRIH responds to the Problem of Consciousness, and how it addresses the two arguments for Primitivism: CA and KA. I conclude that NRIH is a plausible account of the metaphysical status of phenomenal consciousness, and a viable response to the Problem of Consciousness.

SECTION 1

A HYBRID ACCOUNT OF CONSCIOUSNESS

The view that the so-called Problem of Consciousness is not a singular problem, but rather an amalgam of problems, is acknowledged by many.[1] It is reasonable to hold that searching for a single solution to the complex Problem of Consciousness is misguided. As I have already argued, the apparent metaphysical gap between the physical and the phenomenal is underwritten by *two* explanatory obstacles: the –trinsicality gap and the –tivity gap. These two gaps are connected intimately, but must ultimately be regarded as distinct challenges to Physicalism.

We have seen that neither EV nor a Representationalist strategy are capable of dealing with both problems single-handedly. This should not be much of a surprise. If the two gaps are distinct, it would be odd for *both* gaps to be a result of our limited conception of the physical, or for *both* gaps to be bound to the representational status of consciousness. If the claim that we have two distinct problems on our hands is plausible, which it is, then we should expect them to have distinct solutions.[2] In other words, we should adopt a *divide and conquer* strategy of splitting the epistemic gap in two then confronting each half separately.

The first step in implementing this strategy is to clarify the scope of the two components of the hybrid account. Self-Representationalism is a proposal about only the *subjective character* of phenomenal states. Against Kriegel, we are *not* offering a straight-forward representational account of qualitative character. Similarly, RIH now pertains only to the *qualitative character* of phenomenal states. Against the bolder form of this hypothesis evaluated in Chapter 4, we are *not* attributing inscrutables an integral role in the explanation of subjectivity. Combining the two gives us the following core thesis:

---

[1] See Metzinger (1995, p.7) and Van Gulick (2004, p.91).

[2] I discuss this further in Section 3.2 where I suggest that the two apparent explanatory obstacles have a common source.

> ***Core Thesis:*** A mental state is a phenomenal state in virtue of suitably representing itself, and is the type of phenomenal state it is in virtue of the unconceived inscrutable properties that implement it.

I call this account the Neo-Russellian Ignorance Hypothesis (NRIH).[3] Interestingly, as I noted in Chapter 4, Russell himself held that inscrutables were integral to qualitative character but *not* to subjective character. As such, it seems fitting to regard this hybrid proposal as a *neo-Russellian* position. Of course, Russell did not hold anything like a Self-Representationalist theory of subjectivity, but NRIH reflects the spirit of his position.

Even before we extend NRIH beyond the Core Thesis, we have reason to believe that it can undermine the two conceptual gaps. Phenomenal states, qua their subjective nature, *seem* to be onticly distinct from physical states because the self-representational nature of phenomenal states generates the illusion that they are distinct from any objective state. Furthermore, phenomenal states, qua their qualitative nature, *seem* to be onticly distinct from physical states because our limited conception of the physical leaves us conceptually ignorant of the intrinsic physical properties essential to their explanation. Both appearances, however, are deceptive. These central claims of NRIH look very promising, but there remains some work to be done.

If NRIH is to be taken seriously, clearly it must provide *something* more than the Core Thesis. But how much more? Obviously we are under no obligation to reveal the proposed inscrutable properties and offer a theory of qualitative character, nor to give the full physical story of how self-representational states come about. The epistemic gap is the appearance that it is *impossible* to account for phenomenal states in physical terms. NRIH has the resources with which to undermine that appearance of impossibility without having to provide a full theory of the phenomenal. The explanation of consciousness is beyond the scope of the problem at hand, plausibly beyond the proper limits of philosophy and, in light of our proposed ignorance, beyond the reach of our current conceptual repertoire. As such, it would be inappropriate to ask for *too much* detail from NRIH.

---

[3] For pronunciation, think ‘Henry’ without the ‘H’.

Nevertheless, there is a key respect in which NRIH *does* need to give us something more than the Core Thesis. There must be a *plausible integration* of the two components of the account. Inscrutables do not just need to be plausibly responsible for qualitative character – they need to be plausibly responsible for the qualitative character of *self-representational states*. Similarly, the Self-Representationalist account of phenomenal states now has to tell a story that attributes an appropriate role to the *inscrutable intrinsic properties* involved in the implementation of that representation. The explanatory goal-posts have been moved a little, and we must acknowledge the possibility that the plausibility of the two component positions does not survive this shift.

How do we go about showing that the two components can be integrated without indulging in unwarranted speculation? I propose a negative strategy. In the next section, I will identify a series of *problems* for NRIH. These problems take the form of potential new conceptual gaps. The threat is that despite facing up to the –trinsicality and –tivity gaps, NRIH opens up *new* gaps, each of which indicates that it is impossible to get conscious experiences out of the proposed physical base. Driven by these threats, we can start to flesh out the metaphysical proposal in a way that shows how those apparent gaps can be overcome. Effectively, I will be aiming to provide enough detail to capture the logical space of plausibly *possible* physical explanations of consciousness without foolishly trying to provide the *actual* physical explanation of consciousness.

SECTION 2

CHALLENGES TO NRIH

The problems I will consider each ask how it is *possible* for conscious experience as we know it to arise from the relevant physical states – that is, from self-representational states implemented by exclusively physical properties, among which are inscrutable intrinsic properties. As we will see, some of these problems are adaptations of

objections raised against competing accounts of consciousness. This will help us to determine whether or not NRIH overcomes the very objections that led us to rule out those competitors. Of course, the plausibility of NRIH depends not just on the negative project of avoiding problems, but on the positive project of offering a harmonious integration of its two elements. Over the course of this section it should emerge that RIH and Self-Representationalism, far from being an arbitrary pairing, are natural partners.

There are five problems in total. The first two are broadly epistemic, and are concerned with NRIH's ability to accommodate our *knowledge* of qualitative character. The third and fourth problems are better characterised as metaphysical, and challenge NRIH's ability to give us structured qualitative experience from an explanatory base with a divergent structure and a divergent intrinsic nature. The fifth and final problem has metaphysical and epistemic elements, but is fundamentally a practical problem. It suggests that if phenomenal states were as NRIH describes them, they would have no utility, so there is no plausible account of why they would come about. I conclude that all five problems, though worthy of discussion, do not constitute serious objections to NRIH.

2.1. THE RECEPTIVITY PROBLEM

2.1.1. The Problem

In conscious experience, we are aware of the qualitative character of our experience. As such, we have some kind of knowledge of absolutely intrinsic properties such as red-qualities, pain-qualities and sweet-qualities. Our having such knowledge is essential to the Problem of Consciousness as we have formulated it. If we did *not* have such knowledge, we would have no justification for talking of the intrinsic qualities of experience. Without the intrinsic qualities of experience, there would be no –trinsicality gap. Without the –trinsicality gap, the epistemic gap would be much less formidable, and the Russellian Ignorance Hypothesis would be of little value. In summation, whatever else we say about consciousness, NRIH had better be

compatible with our having knowledge of the intrinsic qualities that characterise conscious experience.

When evaluating RIH, I defended an argument for the existence of inscrutables (Chapter 4, Section 1). Within NRIH, those inscrutables are now claimed to be integral to the explanation of the intrinsic qualities of experience. The appeal to inscrutables is only justified if we have a persuasive argument for their existence isolated from any concerns about consciousness.[4] A key premise in that argument was an epistemological thesis: the *receptivity* of human knowledge. This is the claim that we have knowledge of things only insofar as they affect us, and it is this feature of our knowledge that renders the absolutely intrinsic properties of physical objects epistemically inaccessible.

The problem for NRIH is this: how could our knowledge of the intrinsic properties that characterise conscious experience be compatible with Receptivity? If we know the properties of our mental states only through how they affect us, we should not have epistemic access to any intrinsic properties involved in that state. Yet we clearly *do* have such knowledge in qualitative awareness, and our having such knowledge is an integral premise of NRIH. There is thus a tension within NRIH.

We have already seen that Receptivity, and its implications for our access to the intrinsic properties of objects, were central themes in Kant's epistemology. Interestingly, Kant held that what goes for knowledge of external objects also goes for knowledge of ourselves. He argues:

> If…we admit that we know objects only in so far as we are externally affected, we must also recognise, as regards inner sense, that by means of it we intuit ourselves only as we are inwardly affected by ourselves. (B156)

Any intrinsic properties involved in mental states are thus unknowable, since the mind '…intuits itself, as it is affected by itself, therefore as it appears to itself, not as it is.' (B69) If Receptivity entails that we have no transparent grasp of the absolutely intrinsic properties of external objects, it entails the same ignorance of the absolutely intrinsic properties of experiential states (see Van Cleve 2002, p.232). If this line of thought is

[4] This is captured by the Integration Condition on EV.

correct, NRIH is committed to denying the possibility of the very knowledge of qualitative character on which it is founded. We can summarise this problem as follows:

> *The Receptivity Problem (RP):*
>
> RP1) Conscious states are states in which we have knowledge of absolutely intrinsic properties.
>
> RP2) If we have any knowledge of absolutely intrinsic properties, Receptivity is false.
>
> RP3) If Receptivity is false, NRIH is not defensible.
>
> RP4) Therefore, NRIH is not defensible.

I previously indicated that the problems explored in this section will each take the form of a conceptual gap between NRIH's proposed explanatory base and phenomenal consciousness. The argument above does not look like it takes the form of a conceptual gap, but it can easily be translated into such terms: the proposed explanatory base involves intrinsic properties to which we have no epistemic access, the explanandum involves intrinsic properties to which we have intimate epistemic access, and there is no possible explanation of known intrinsic properties in terms of unknowable intrinsic properties.

2.1.2. Response

Where is the weak point in this argument? The inference to RP4 is clearly valid. RP1, as discussed, is hard to deny and is integral to NRIH. RP3 is difficult to resist – we cannot give an argument for inscrutables without recourse to Receptivity, and we cannot defend NRIH without an argument for inscrutables. That leaves RP2. Perhaps we could argue that Receptivity is compatible with knowledge of intrinsic properties. In that case, however, the argument for inscrutables would surely collapse.

Alternatively, we could argue for a *qualified* version of Receptivity. Most of the time our epistemic situation leaves us ignorant of the intrinsic properties of things, but the knowledge we gain in conscious experience is an exception to the rule. The intrinsic properties of external objects are inscrutable, but the intrinsic properties of internal experience are not. This qualified claim is perfectly compatible with the

defensibility of NRIH: it is specifically our conceptual ignorance of *non-experiential* intrinsic properties that is held to make experiential intrinsic properties appear physically inexplicable. However, this qualified version of Receptivity looks *ad hoc.* How could we possibly justify the claim that the knowledge we gain in phenomenal awareness is an exception to the unknowability of intrinsic properties? I suggest that the solution to the Receptivity Problem is indeed to adopt a qualified version of Receptivity, but that this proposal need not be *ad hoc*. On the contrary, NRIH *predicts* that we can have epistemic access to absolutely intrinsic properties in conscious experience, and that this is the *only* possible context in which we can have such epistemic access.

Van Cleve seeks to challenge the Kantian view of self-knowledge. He suggests that though we know external objects via a causal chain from the object to ourselves, if we could have knowledge of the *terminus* of a causal chain, we would not be limited to knowledge of causal powers. He concludes that '…there can be no argument against the possibility of acquaintance sheerly from the causal nature of perception.' (2002, p.229) The idea is that if there is a kind of knowledge that is *not* causally mediated, there is no reason why such knowledge could not disclose intrinsic properties rather than merely causal dispositions.

This opens up a promising logical space, but does not finish the job. Why should it be the case that knowledge of qualitative character is a case of unmediated epistemic access? Where we represent external objects, there is causal mediation. Where we represent that representation of the world, there is plausibly a distinct mental state representing the first, and so another case of causal mediation. The chain can continue with us having further higher-order mental states representing lower-order states, but at every step a state is known only through how it affects some distinct state. Being 'in the head' does not mean we can magically have knowledge of something without causal mediation. The Receptivity of human knowledge plausibly still applies. The challenge is thus to understand how consciousness could possibly provide epistemic access without causal mediation.

Fortunately, Self-Representationalism provides us with precisely what we need. We have considered strong arguments for the conclusion that all phenomenally

conscious states are self-representing states (Chapter 5, Section 3). We have also considered the distinctive epistemic situation that such self-representation generates. As Kriegel proposes, '…knowledge of consciousness, and of consciousness alone, does not require causal contact with the known.' (2009, p.295)[5] Causal contact is required if a mental state is to constitute knowledge of some state or object *distinct from it*. Receptivity thus applies to all cases in which the represented object of knowledge is distinct from the representing 'knowing' state. In self-representation, however, the representation and the represented are *identical*. As such, causal contact with our conscious states is not necessary for knowledge of our conscious states. Consequently, there is no reason to deny that in consciousness we can have transparent knowledge of *intrinsic* properties. The qualitative character of a conscious experience is epistemically accessible precisely because the state with that character and the representation of that state are one and the same state.

We do not yet have a full account of how such knowledge works, nor of the metaphysical status of the phenomenal qualities, but the self-representational nature of conscious states provides NRIH with a satisfactory way out of the Receptivity Problem. Importantly, we have not had to contrive any implausible *ad hoc* epistemic story in order to achieve this. The non-causal nature of our knowledge of consciousness was built into Self-Representationalism from the outset, and preceded any attempt to combine it with EV. This is a clear case of the two components of NRIH *complementing* one another rather than being an awkward conjunction. Self-Representationalism solves the mystery of how we could have knowledge of qualitative character.

The response we have offered to the Receptivity Problem has an interesting relationship with certain historical positions. As we have seen, the Receptivity of self-knowledge was advocated by Kant. Some followers of Kant, however, agreed that our knowledge of *external* objects is mediated, with all the epistemic limitations that entails, but proposed that our *self*-knowledge is of a very different kind. Schopenhauer famously proposes that '…a way *from within* stands open to us to that real inner

[5] The epistemic implications of target states being 'embedded' or 'contained' within the state that represents them have been explored by a number of figures. For instance, Burge (1988) argues that such containment is responsible for the distinctive epistemic authority of self-knowledge.

nature of things to which we cannot penetrate *from without*' (1819/1844, 2.195). The metaphor of knowledge from within and knowledge from without is intriguing and has been echoed by many other thinkers. However, this metaphor must be cashed out in a rigorous way if it is to be taken seriously. NRIH can do just this. Knowledge 'from without' is the epistemic contact a subject has with entities distinct from themselves through how those entities affect them. Knowledge 'from within' is the reflexive non-causal epistemic contact provided by self-representational mental states.

It is interesting that Schopenhauer's manoeuvre here leads him close to panpsychism: a view into which NRIH must not collapse. The driving thought seems to be that our 'inside' knowledge must be analogous to that 'inner' nature of external objects which perception fails to penetrate. Since our 'inner' nature involves awareness, so too does that of inanimate objects. NRIH maintains the spirit of this thought whilst avoiding any threat of panpsychism. The intrinsic properties of experience are in some way bound to the intrinsic properties of fundamental physical entities. Exactly how close that association is we will discuss later, but since NRIH claims that phenomenal qualities are the upshot of inscrutables, this surely gives us some kind of epistemic contact with intrinsic physical properties. Panpsychism, however, is avoided since it is not held that all physical entities have the kind of internal awareness that we have. Remember, according to NRIH our 'inside view' only exists thanks to our complex self-representational mental architecture. The external objects we perceive do not generally have such self-representational states, so we need not attribute them anything even analogous to consciousness. Everything has an 'inner nature' constituted by their intrinsic properties, but only minded creatures like us have an 'inner awareness'.

The Receptivity Problem constitutes a *prima facie* tension within NRIH, but it is clear that the position has the resources with which to overcome it. Furthermore, the suggested solution reveals how the components of NRIH complement each other: a problem associated with the EV component was solved thanks to the Self-Representationalist component. The proposal taps into a line of thought that goes back at least to Schopenhauer, but cashes out an idea that was previously expressed only as metaphor, and avoids the unwanted threat of panpsychism.

2.2. THE CONTENT PROBLEM

2.2.1. The Problem

The Receptivity Problem is based on epistemic concerns. NRIH does not just have to accommodate the *existence* of qualitative character; it must also accommodate our *awareness* of qualitative character. After all, what makes a quality a *phenomenal* quality is that it characterises a state of subjective awareness. The previous problem threatened to show that NRIH was incompatible with such awareness. Though that problem was overcome, there is a further problem that presents a similar threat. A consideration of the nature of self-representational states suggests that it is impossible to get awareness of intrinsic qualities from the proposed explanatory base. The worry here is not to do with the Receptivity of knowledge, but with the role of inscrutables in phenomenal representations.

According to Self-Representationalism, phenomenal states involve two layers of content. There is a first-order layer that represents features of the world beyond us, and a second-order layer that represents that very representation. NRIH adds something to the story here by suggesting that the physical states that implement phenomenal representations involve intrinsic properties of which we have no conception – that the vehicle of a phenomenal representation includes inscrutable properties. But how could introducing intrinsic properties into the vehicle of representation account for our *awareness* of phenomenal qualities? In representation we are, at best, aware of the content of that representation, not of the properties that implement it. We can summarise the problem as follows:

> *The Content Problem (CP):*
>
> CP1) By being in a self-representational mental state M, we are aware only of the content of M.
>
> CP2) Inscrutables have a vehicular role in M, but are not responsible for the content of M.
>
> CP3) Therefore, inscrutables are not responsible for what we are aware of by being in M.

CP4) In M, we are aware of phenomenal qualities.

CP5) Therefore inscrutables are not responsible for the phenomenal qualities that characterise M.

I use the phrase 'responsible for' as a minimal commitment of NRIH. Exactly what the relationship between inscrutables and phenomenal qualities is needs further discussion, but *whatever* it is, a state had better have its qualitative character in virtue of the inscrutables involved in its implementation. That much is entailed by the Core Thesis of NRIH.

2.2.2. Response

I will ultimately challenge CP2, but first it is worth noting a potential challenge to CP1. Perhaps it is not the case that everything of which we are aware in virtue of M is part of the content of M. In Chapter 5 I mentioned *impure* intentional properties. These are *ways* of representing that can vary between states with the same content. It could be held that qualitative character is fixed not just by the content of a conscious state, but by its *manner* of representation. If the inscrutable properties that implement M make a difference to the impure intentional properties of M, the path may be open for us to have some kind of awareness of inscrutables or, more precisely, have a state of awareness to which M's inscrutables make a phenomenally manifest difference.

There are at least two reasons to disregard this line of thought. One, the 'adverbial' view of qualitative character it proposes is deeply implausible (as discussed in Chapter 5, Section 2.2.2). Two, even if qualitative character *was* determined by impure intentional properties, it is unclear why and how inscrutables would be responsible for those impure properties. Of all the vehicular properties that carry a phenomenal representation, why would it be inscrutables that determine the manner of representation, and how would their doing so explain the qualitative character of experience?

A more promising route for NRIH is to challenge CP2. Perhaps the inscrutable properties that contribute to M do not only perform a vehicular role. Perhaps they are part of M's content. That is, our phenomenal representations represent their own intrinsic physical properties. M represents both properties in the external world and

properties of itself. After all, NRIH already claims that M is self-representing, so why not say that it represents its own intrinsic properties? As Lockwood proposes, '[w]hy should one not think of awareness precisely as *disclosing* certain intrinsic attributes of states of…our brains…?' (1989, p.162) The qualitative redness of a phenomenal state would then be both a constituent of that state and something represented by that state: both vehicle and content. Of course, this raises questions about how a state could instantiate qualitative redness in virtue of its inscrutables. I will address this concern in the next section, but here we at least have a promising proposal about the *representational status* of inscrutables.

A potential problem for this proposal is that it fails to respect the phenomenology of qualitative awareness. It is not the case that we are aware of the world and separately aware of properties of the representing state. Experience does not have this two-level structure. We do not have a first-order component that represents the colour of the post-box, and a higher-order component that represents a red-quality of that representation. Instead, we are transparently aware of the redness as a property of the post-box. The redness we find in experience is indeed part of the content of our experience, but it is not represented as a feature of our mental states, it is represented as a feature of the external world. As such, the proposal that M represents its own inscrutable properties does not give a defensible account of the qualitative character of experience.

There is a way out of this problem for NRIH. In line with CP1, phenomenal qualities had better be *represented* by phenomenal states, for how else could we be aware of them? In line with our phenomenology, those qualities must be represented in a way that accounts for their appearing as external properties of the world, rather than internal features of our representations of the world. To reconcile these two commitments, the approach I have in mind is *projectivist*. Projectivism is the view that certain features attributed to the world by our mental representations are actually features of our selves. Our representations *project* their own properties on to the world that they represent.

In a sense, M represents its own qualities, which it has in virtue of the inscrutable properties involved in its implementation. However, it represents them *as*

properties of external objects. The qualities are thus part of the content of M, and M's capacity to represent them is bound up with its self-representational nature. But we are not thereby committed to a two-level phenomenology. The perceived red-quality of the post-box is, in an important sense, a property instantiated by a state of our brain. We are aware of it because we represent it, but we are aware of it *as* an external quality because we represent it *as* an external quality. There is nothing incoherent about suggesting that our access to phenomenal qualities is misleading in this way. It might be counter-intuitive and it might need further explanation, but that should not worry us.[6] The key point is that inscrutables are given a status in representation that is not at odds with what our experience of phenomenal qualities is like. To overcome the Content Problem I have rejected CP2. Inscrutables are not just vehicular properties, they are represented by our phenomenal states. However, the *way* in which those properties are represented is such that they are projected on to the external objects of experience, rather than being represented as features of the representation itself.

When responding to the Receptivity Problem I cited some historical precedent to illuminate my proposal. We can do the same here. The idea of projection has a strong philosophical history. It is plausibly integral to Locke's familiar notion of secondary properties (see Egan 2010). Hume proposes that '…the mind has a great propensity to spread itself on external objects…' (quoted Egan 2010, p.69). Kant, to whom we have turned at many points, can also be understood as advocating a kind of Projectivism.[7] More recently, Lockwood has presented a position not dissimilar to NRIH. He suggests that phenomenal qualities are instantiated within the brain, but are 'taken as' external properties by our representations. He argues that '…there is no consciousness, no sentience, without *taking as*.' (1989, p.312)[8] Since any awareness of

[6] For some useful work on why and how we project properties see Jakab (2003). This paper includes responses to some prominent objections to Projectivism.

[7] Kant's examination of the self-world duality in the Transcendental Deduction could be taken to show that it is only possible to experience something if it is presented as *distinct* from the self. As such, the notion of experiencing phenomenal qualities *as* features of one's representations is deemed impossible. A kind of projection is thus necessary to their being experienced at all.

[8] Note, Lockwood's position is a form of Type-F Monism as discussed in Chapter 4 (Section 4.1). Unlike Lockwood though, NRIH does *not* regard qualities as onticly primitive.

qualities must involve taking them *as* something or other, their being taken as external properties is no more mysterious than their being taken as features of the brain.[9]

More recently still, the notion that conscious experience is a 'virtual reality' has gained respect among cognitive psychologists and their philosophical allies.[10] The idea is that our mind constructs a kind of model of reality, but this virtual model is *taken as real*. Our awareness of the world is mediated by an awareness of this mental model, but we take this to be direct access to the external world. The same idea can be applied to our awareness of phenomenal qualities: they are internal properties taken as external properties. Obviously there is a lot more to be said about how such projection works and why it would occur, but we have come far enough to undermine the Content Problem. Inscrutables can be integral to the content of M, so are suitably positioned to be responsible for the qualitative character of M.

2.3. THE QUALITATIVE CHARACTER PROBLEM

2.3.1. The Problem

Our responses to the previous two problems have helped us to flesh out NRIH. The proposal is now that phenomenal qualities are the result of the inscrutable intrinsic properties that implement our phenomenal representations, that those qualities are *epistemically accessible* in virtue of the self-representational structure of those states, and that the representation of those qualities involves a *projection* of them onto the external world. The next problem is perhaps the most serious, and is one around which I have been circling. In the original discussion of RIH, I was keen to emphasise that inscrutables are not simply unexperienced phenomenal qualities (Chapter 4, Section 4.1). It is not the case that the dispositions of entities such as electrons are grounded in qualities such as redness. The question is this: how can M's access to its own inscrutable properties give rise to an experience of qualitative redness if those properties are not themselves red-qualities? According to NRIH we have no concepts

[9] See Boghossian & Velleman (1991), Wright (2003) and Coates (2009). The current proposal also has echoes of indirect realism and sense-datum theory.
[10] See Metzinger (2003, Ch.8).

with which to characterise inscrutables. Nevertheless, the account makes a claim about what inscrutables are *not*. They are not instantiations of the qualities with which we are familiar in conscious experience. To say that inscrutables have a *non-red* nature may be enough to cast doubt on the possibility of qualitative redness arising from inscrutables. The same worry applies to *all* phenomenal qualities.

Perhaps NRIH can just re-apply EV's main strategy here: it only *appears* that you cannot get red-qualities out of non-red intrinsic properties because we have a *limited conception of non-red properties*. Neural properties are indeed non-red properties from which you can never get qualitative redness, but this is because they are of the wrong metaphysical category: these *structural* properties cannot be responsible for *intrinsic* properties like redness. Inscrutables, by contrast, are intrinsic properties. We cannot apply what we know about *familiar* non-red properties to *inscrutable* non-red properties. The central thought behind EV is that the properties of which we do have a conception present a skewed picture of the physical world, so we cannot assume that what we know of familiar properties also applies to the unfamiliar ones. Yet this is just what we are doing if we insist that you cannot get red-qualities out of non-red inscrutables.

How might a critic of NRIH respond to this manoeuvre? They would need to be able to show that it is not just a shaky intuition that motivates the claim that non-red inscrutables cannot be responsible for the qualitative redness of an experience. They must appeal to some genuine conceptual gap. I think the most plausible route here is to consider the various possible relations of ontic dependence that might hold between inscrutables and phenomenal qualities, and cast doubt on each of them individually. I will consider identity, composition and implementation.

First, identity can be ruled out immediately. Identity requires that phenomenal qualities and the relevant inscrutables are not different in any respect. Since we are discussing NRIH's commitment to non-red inscrutables being responsible for red-qualities, it is clear that a straight-forward identity relation is not available.

Second, the relation could be one of *composition*. For instance, a wall with the property of being 6m high might have that property thanks to a collection of bricks each with the property of being 3cm high. There is nothing mysterious about how the

higher-level property comes about in virtue of quite different lower-level properties. However, in the case of phenomenal qualities, this kind of model cannot apply. We know from experience that phenomenal qualities are *simple*. They are non-composite, they have no parts, they are not a sum. Consequently, it cannot be the case that red-qualities are *made of* non-red-qualities, because red-qualities are not made of *anything*.

Third, it could be held that inscrutables *implement* phenomenal qualities, where implementation is a relation of ontic dependence in which a higher-level property is realised by a lower-level property. This need not be a part-whole relation, so is distinct from the notion of constitution. The problem for NRIH here is that phenomenal qualities do not look like the kind of properties that can be implemented by anything. The implementation of a higher-level property involves a lower-level property performing a *role*. For instance, where a person performs the role constitutive of being a King, they *implement* King-hood. Phenomenal qualities, however, exceed any characterisation in terms of causal role. In fact, it is precisely their *non-structural* nature that has driven our whole discussion of them. Absolutely intrinsic properties are, by their very nature, non-structural properties, and only structural properties can be implemented. The claim that an instantiation of qualitative redness requires nothing more than the performance of a certain role amounts to *functionalism* about phenomenal qualities, and we have already established that this not a viable proposal. Phenomenal qualities are simply not the kind of property that can be implemented by more basic properties – inscrutable or otherwise – performing a particular role.

Identity, composition and implementation plausibly constitute an exhaustive list of the relations of reductive ontic dependence. We can summarise the problem for NRIH as follows:

> *The Qualitative Character Problem (QCP):*
>
> QCP1) If non-qualitative inscrutables are responsible for phenomenal qualities, they are either identical to phenomenal qualities, constitute phenomenal qualities or implement phenomenal qualities.
>
> QCP2) If non-qualitative inscrutables are identical to phenomenal qualities, then they are not really non-qualitative.

QCP3) Phenomenal qualities are not the kind of property that can be constituted by anything more basic than themselves.

QCP4) Phenomenal qualities are not the kind of property that can be implemented by anything.

QCP5) Therefore non-qualitative inscrutables are not responsible for phenomenal qualities.

2.3.2. Response

How can NRIH escape this problem? QCP2 and QCP4 should be conceded, but both QCP1 and QCP3 can be challenged. Against QCP1, we are not in a position to conclude that identity, composition and implementation are the only relevant relations of ontic dependence. There is the possibility of *unknown* relations of dependence. Consider, for instance, the proposal put forward by McGinn (1994, 2004 Ch.8). McGinn suggests that we are only capable of understanding explanations that involve 'combinatorial atomism with law-like mappings' ('CALM' principles for short). He proposes that the explanation of consciousness involves *non*-CALM explanatory principles beyond our cognitive reach. Without committing to McGinn's specific hypothesis, we can take seriously the general possibility of unknown relations of ontic dependence.

A related proposal is that the relevant form of ontic dependence involves 'fusion'. Two entities 'fuse' when they form a whole but, unlike in cases of composition, the parts do not persist when the whole is formed. Entities have certain propensities such that, in appropriate circumstances, they necessitate the introduction of a new entity with different properties to its predecessors. This entity could not have existed without its predecessors, but it *replaces* them rather than being a mere *aggregate* of them. This notion has been explored in the context of panphenomenalist accounts of consciousness, but could also be used to support NRIH.[11] Perhaps inscrutables 'blend' to form phenomenal qualities. When the right inscrutables come together in the right circumstances, they entail the occurrence of qualitative redness. This quality is a new property with no constituent parts, but its occurrence is entailed a priori by the occurrence of properties that are not red in and of themselves. On this

[11] See Seager (2010) and Coleman (2012) for more on combinatorial infusion.

view, inscrutables have unknown fusion-propensities that are responsible for the occurrence of phenomenal qualities.

This appeal to unknown forms of dependence casts serious doubt on QCP1, but it is a speculation that pushes at the limits of our understanding. It would be preferable to undermine the Qualitative Character Problem in a more straight-forward way. We can do this by challenging QCP3. Perhaps phenomenal qualities are constituted by inscrutables, but the apparently non-composite nature of those qualities is illusory. A red-quality in our experience plainly *seems* to be non-composite, but why infer that it *is* non-composite? According to the picture offered by NRIH, qualitative awareness involves us accessing the inscrutable properties of our representational states. We are not committed to a *full disclosure* of those properties. We have already said that they appear to us as external properties even though they are really internal features of our representation. Here we can suggest that our experience is misleading in a further respect. The qualitative character of our experience appears to involve non-composite qualities, but in reality those qualities each have a number of inscrutable constituents.

Yet again, we can cite some weighty historical precedent to help put this proposal in focus. Leibniz proposes that '…sensible ideas appear simple because they are confused and thus do not provide the mind with any way of making discriminations within what they contain.' (quoted Puryear 2005) 'Confusion' is simply a failure to represent the constituent parts of something, making it appear simple when it is not.[12] The access we have to our own inscrutables in consciousness could thus be a 'confused' kind of access. Leibniz considers the idea of confusion in the context of colours: 'We now have a complete analysis of green into blue and yellow…yet we are quite unable to discern the ideas of blue and yellow within our sensory idea of green, simply because it is a confused idea.' (quoted Puryear, 2005) This claim is disputable, but if apparently simple colour qualities are plausibly composites of other colour qualities, why not take the next step and say that basic

[12] Pereboom (2011) has recently explored the possibility that we misrepresent phenomenal qualities as simple. However, unlike the current proposal, he blames that misrepresentation on our *introspections* of phenomenal states, rather than on the content of the phenomenal state itself.

colour qualities are themselves composites of *non*-colour properties beyond our current comprehension?

It could be objected that deploying the notion of 'confused' epistemic access entails that we are not really aware of our inscrutables at all, which contravenes the response to the Content Problem above. How can it be the case that the properties of which we are phenomenally aware are non-red inscrutables, but for our awareness to be characterised by simple red-qualities? Again, Leibniz can clear things up for us. Leibniz claims that the human mind is made up of countless perceptions, but we experience the overall 'roar' of these parts:

> To hear this noise as we do…we must hear the parts which make up this whole, that is the noise of each wave, although each of these little noises makes itself known only when combined confusedly with all the others, and would not be noticed if the wave which made it were by itself. (quoted Look, 2008)

Though Leibniz is talking about a vast collection of *petites perceptions* constituting our experience, NRIH does not attribute any mental features to inscrutables.[13] Nevertheless, the status of inscrutables could be analogous. The relevant inscrutables could each play a role in the qualitative character of our awareness, even though we cannot make out those properties individually. Since we are relatively high-level entities, there could be an enormous number of inscrutables in play in our qualitative awareness. Just as the noise of the sea can differ greatly from the noise of each drop of water, so too the qualitative upshot of this field of inscrutables could differ greatly from the nature of those inscrutables taken in isolation.

Overall, NRIH can evade the Qualitative Character Problem. It is hard to get our heads round non-red inscrutables being responsible for red phenomenal qualities, but this much is to be expected given our proposed limited epistemic position. The relation between inscrutables and phenomenal qualities might itself be a kind of relation beyond our current conception, perhaps involving something like 'fusion'. Alternatively, if we make the Leibnizian move of regarding the apparent simplicity of phenomenal qualities as a case of 'confusion', the relation could be one of

[13] As with our discussion of Schopenhauer in Section 2.1, NRIH is drawing on a historical predecessor but purging it of its panpsychist elements.

composition. I will stick with the Leibnizian move, though the alternative response can be kept in reserve.

2.4. THE STRUCTURAL DIVERGENCE PROBLEM

2.4.1. The Problem

The inscrutable properties instantiated within our brain will have certain structural features such as their spatial distribution. The phenomenal qualities that characterise our conscious experience also have certain structural features, including their distribution in our experiential field. The problem for NRIH is that the structure of inscrutables and the structure of phenomenal qualities is likely to diverge, and it therefore seems impossible to account for the latter in terms of the former.

This has been raised as an objection to panphenomenalist forms of Type-F Monism, so it is important to show that NRIH is not vulnerable to the same objection. The objection to panphenomenalism is that the micro-experiential entities claimed to be responsible for human consciousness have structural features that diverge from those of the conscious states they are supposed to constitute. Even though NRIH makes no appeal to micro-experiential entities, the same structural divergences plausibly occur.

Unlike the previous objection, this problem is not concerned with whether the intrinsic nature of inscrutables could be of explanatory relevance to phenomenal qualities. It is simply concerned with the *structural distribution* of inscrutables, which is well within our epistemic grasp. An appeal to the unknown nature of inscrutables will not help NRIH when it is their structural status that is causing trouble. We can distinguish between three types of structural divergence in ascending order of seriousness:

A) The Configuration Problem: Consider a phenomenal experience of a French flag. The colour-qualities that characterise that experience have a certain structure. If

inscrutables are responsible for those qualities, they had better have the same structure. After all, if they do not have that structure, we are left with the mystery of how it is possible for properties with one structure to give rise to experiences with a totally different structure. But it is implausible that inscrutables have the same structure as the qualities of that experience. Inscrutables are the absolutely intrinsic properties of fundamental physical entities. As we move around to get a different perspective on the flag, is it really plausible that there are corresponding movements of those fundamental entities in our brain? Is it plausible that when we open our eyes, these entities spring into action and jump into a configuration isomorphic to what we see? Clearly not.

B) The Grain Problem: A further problem is that our experiences present us with a smooth qualitative field. However, the inscrutable properties claimed to be responsible for the field are discrete, since the entities that have those intrinsic properties are discrete. Even if they had a configuration close to that of our qualitative field, the inscrutables would be 'gappy' but our experience would involve a smooth qualitative field.[14] How is it possible for discrete inscrutables to generate smooth qualitative fields? If it is an inexplicable mystery, NRIH fares no better than its competitors.

C) The Combination Problem: This final problem is subtly different to the second. Even if inscrutables were configured in the same way as our qualitative field, there is a deeper respect in which they will still diverge structurally. Our phenomenal states have a natural *unity*. The various phenomenal qualities that contribute to our experience form a single experiential space, and the qualities within that space stand in a special relation to each other that they do not stand in to qualities of someone else's experience. If we lined up two people, their respective experiences would remain distinct. If we somehow squished their brains together so that the relevant fundamental physical entities touched, we would still have distinct experiences, or at

[14] See Lockwood (1993), and Stoljar's (2001) discussion of Maxwell.

least have no explanation for why a single experience should come about. There is more to the unity of a qualitative field than mere spatial contiguity. If a dozen inscrutable properties are configured side by side within us, with no gaps between them, why don't we get a collection of twelve qualitatively simple experiences? What accounts for their unification in a single qualitative field?[15]

The schematic form of this problem for NRIH can be captured as follows, and each of the three variations above can be used to fill in the place-holders:

> *The Structural Divergence Problem (SDP):*
>
> SDP1) A subject's inscrutables have a certain structure *X*.
>
> SDP2) A subject's qualitative field has a certain structure *Y*.
>
> SDP3) *X* and *Y* are divergent.
>
> SDP4) There is no accounting for properties with one structure in terms of properties with a divergent structure.
>
> SDP5) Therefore there is no accounting for a subject's qualitative field in terms of their inscrutables.

2.4.2. Response

The ways in which we have already fleshed out NRIH are such that this problem should not seem too worrying. Nevertheless, it is worth working out exactly how it should be dealt with. The first option is to deny SDP3. Indeed, many panphenomenalists and panqualitativists make just this move. Lockwood (1993), for instance, suggests that on the quantum level our brain states form unified quantum fields. These fields are one-and-the-same as our fields of experience, so there is no structural mis-match.[16] The strategy is simply to maintain that inscrutables have the same structure as our phenomenal states. This move can be made on the level of configuration, grain and unity. This may seem implausible, but it is far from impossible. There is thus no knock-down criticism of NRIH here. Nevertheless, I think we can present a more palatable solution to the problem.

---

[15] This is adapted from William James's (1890) famous objection to panpsychism. For a more recent version of the objection, see Goff (2006).

[16] Related moves are made by Seager (1995/2010) and by Coleman (2012).

We can simply deny SDP4. It is not the case that inscrutables must have the same structure as our qualitative field in order to be of explanatory relevance. That is far too simplistic a view of the proposed explanatory relation. Our experiences would not have the qualitative features they have without inscrutables, but inscrutables need not be solely responsible for all aspects of our experience, and need not constitute our experience like pixels on a screen.[17] We have already shown how the representational features of a phenomenal state play an integral role in determining our phenomenology. Just as internal properties are represented as external, and a complex of properties is confusedly represented as simple, so too a set of properties with one structure can be represented as having some divergent structure. Our representations *bestow* structure on our experience. It shapes them and unifies them in a single qualitative field. A smooth qualitative field, for instance, need not entail the existence of anything literally smooth in the head. Things *seeming* some way to us does not entail the occurrence of anything that literally *is* that way.

If our phenomenology can diverge so far from the physical state that grounds it, why not just scrap the appeal to inscrutables? Our brain state could simply *represent* the occurrence of intrinsic phenomenal qualities even though nothing like them is actually instantiated. However, as discussed in Chapter 5 (Section 2.2.1), this move is not open to the Representationalist. On a standard Physicalist ontology, there are no intrinsic properties to be represented, and we cannot make sense of brain states coming to represent properties that are nowhere instantiated. Intrinsic properties need to figure in the story somewhere, and NRIH is an attempt to sketch the most plausible place for them. Representational factors have been attributed a key role in NRIH's account of consciousness, but only insofar as they do not threaten to compromise Physicalism.

Keeping up our theme of establishing historical precedent, NRIH now seems to be following Kant in distinguishing the *form* and the *matter* of experience. There are two components to experience: the chaotic manifold that we get through being affected by the world, and the order that the mind imposes upon that data. Inscrutables have a status reminiscent of the Kantian 'manifold' of sensation. Our self-

[17] Stoljar (2001) offers this response to the 'grain problem' in his defence of RIH.

representational states provide the structure of our experience: they give our subjective reality its spatial configuration, its unity, and its apparent externality. But all that structure would amount to nothing to the subject without some substance to fill it in, and this is just what inscrutables provide. The specific nature of the inscrutables determines the specific qualitative character of each point in our experiential structure.

The responses to all four of the problems described so far involve letting the representational features of our conscious states do some of the heavy lifting. The original motivation behind NRIH was to help EV out by letting Self-Representationalism deal with the –tivity gap. It now emerges that the representational features of conscious states can allow us to close all the smaller gaps that threaten to derail NRIH as well. This most recent suggestion also indicates how Self-Representationalism can enjoy some benefits from the union. The problem for standard forms of Representationalism is that they can plausibly account for the structure of experience, but not its non-structural features: for its form, but not its matter. Inscrutables are introduced to give phenomenal representation its matter. Without them, our representations would be merely structural, and could never constitute a phenomenal reality for their subject. The pairing of RIH and Self-Representationalism is looking promising.

2.5. THE PURPOSE PROBLEM

2.5.1. The Problem

The four problems discussed have each asked *how could conscious experiences arise from the proposed explanatory base?* I have developed NRIH's explanatory story in a manner that should assuage those worries. Our final problem is of a different kind. It asks *why would the proposed explanatory base occur?* Even if it is allowed that the right self-representational and inscrutable properties are sufficient for conscious experience, we are owed an account of why it is that such states come about. What *purpose*, according to NRIH, could phenomenal states have? We will see that certain

considerations indicate that representing our own inscrutable properties achieves nothing. The problem is that if this achieves nothing, serious doubt is cast on NRIH. Our brain is an efficient machine that does things that serve our functional economy and help us as organisms. This is bound to the fact that our brains are the product of an evolutionary process. With a few exceptions, our characteristics aid our survival, especially when they are complex traits, so we should expect consciousness to have some evolutionary value. As such, if NRIH is committed to qualitative awareness being a useless side-show, its plausibility would be dramatically diminished.

Why is there any threat of NRIH entailing that qualitative awareness is purposeless? It comes down to causal considerations, and a revival of objections previously wielded against Primitivism. The Primitivist view that phenomenal properties are onticly distinct from physical properties entails that for every one of our physical actions, we would still have performed that action in the absence of our phenomenal states. There are two layers to this: we would have done the same thing had we had *no conscious awareness at all*. That is, our zombie twin would have the same functional profile as us. Second, we would have done the same thing had our conscious awareness had a *different qualitative character*. That is, our invert twin, for instance, would have the same functional profile as us. In Chapter 1 I argued that these were both unacceptable commitments. When you say 'I am conscious' this should be caused by your being conscious, and when you say 'Ouch!' this should (at least sometimes) be caused by your experience having the particular qualitative character it has.

To what extent does NRIH face the same problems? The good news is that self-representational states are an efficacious part of our functional economy. Even if we do not know precisely what function they serve, it is plausible that self-representation achieves *something*. Representing our own mental states plausibly changes our functional profile in some way.[18] This is the kind of thing that an efficient machine like the brain could plausibly spend its resources on, and which could have given our

[18] This is discussed by Kriegel (2009, Ch.7). There is also the possibility of drawing on claims made by various HOR theorists in this area. The fact that we have a rich variety of plausible options is enough to assuage doubts about the possible utility of self-representational states. As such, I will not advocate any specific account.

distant ancestors the kind of evolutionary advantage that accounts for the place of self-representational states in humans today. This means that the *subjective awareness* side of consciousness does not raise worries of purposelessness for NRIH. What of the *qualitative character* side? I will propose variations on the invert and zombie thought-experiments which indicate that NRIH is committed to qualitative character serving no purpose.

NRIH claims that the qualitative character of experience depends on the specific nature of our inscrutables. Presumably then, mixing up your inscrutables would mix up the qualities of your experience. For instance, whatever inscrutables are responsible for colour qualities could be reversed to produce a being whose colour-qualia are inverted relative to our own. Since such an inversion need not entail any *functional* difference, it is possible for a being to have precisely the same functional profile as us, but to differ from us with respect to their qualitative character. Call such a being an 'inscrutable-invert'. Note, such a being is not a complete *physical* replica of us since they differ from us with respect to their intrinsic physical properties. Nevertheless, the possibility of such a being would have troubling ramifications for NRIH.

If my inscrutable-invert twin has the same functional profile as me, then the specific qualitative character of my experience never contributes to my behaviour. I would have done the same thing had my inscrutables been such as to generate an entirely different experience. Perhaps this is not too worrying in the case of our colour qualities being re-arranged, but the implication is that *all* of our phenomenal qualities could be mixed up without making a difference to what we do. The question for NRIH is this: if the qualitative character of our conscious states makes no difference to our functional profile, why would our brains conspire to bring about qualitative awareness? If differences in qualitative character are not a difference that *makes a difference*, qualitative awareness is rendered pointless.

Can a parallel objection to NRIH be developed involving zombies rather than inverts? Perhaps NRIH is committed to the possibility of beings whose representational states have *no* qualitative character, but who can do all the same things we do. A first pass at showing that NRIH has this unfortunate commitment is to hold that it is

possible for a being to have the same functional profile as us, but to lack inscrutables. Since NRIH claims that inscrutables are essential to qualitative character, such a being would be a zombie. As such, NRIH allows the possibility of zombies with the same functional profile as us, so renders qualitative character pointless.[19] However, this initial line of argument clearly fails. According to NRIH, inscrutables are necessary for the instantiation of any causal properties. A twin with the same functional profile as you but without inscrutables is thus incoherent: a being cannot have the same functional profile as you without absolutely intrinsic properties. In fact, it cannot even *exist* without such properties. Consequently, the standard zombie scenario does not threaten NRIH.

However, NRIH cannot escape that easily. In light of our previous conclusions, the mere *possession* of inscrutables by our self-representational states is not sufficient for qualitative awareness. Those inscrutables have to be suitably *represented*. Consequently, it seems possible for there to be a being like us on the level of inscrutables, but who does not *represent* those inscrutables, and so does not have qualitative awareness. Call such a being a 'pseudo-zombie'. The proposed representational difference probably entails some functional difference. The problem is that this difference is unlikely to be a significant one. Surely it is not integral to the way we process stimuli and govern behaviour that we represent our inscrutables along the way? Our unconscious states do a perfectly good job without such representation, and our self-representational states could achieve whatever extra it is they achieve without representing inscrutables. Why don't we just represent the structural properties of the world, and of our own mental states, and never represent intrinsic properties? What advantage does representing our inscrutables give us over the pseudo-zombie? If NRIH cannot answer these questions, serious doubt is cast on its potential to attribute an appropriate causal role to qualitative awareness:

> *The Purpose Problem (PP):*
>
> PP1) If NRIH is true, qualitative awareness makes no functional difference to its subject (or only makes a negligible functional difference).

[19] Again, this being is our *functional* duplicate rather than our *physical* duplicate, so the objection is not directed at showing that consciousness involves non-physical properties.

PP2) If qualitative awareness makes no functional difference to its subject (or only a negligible functional difference) it has no purpose in our mental economy, and no plausible evolutionary origin.

PP3) Qualitative awareness must have a purpose in our mental economy and a plausible evolutionary origin.

PP4) Therefore NRIH is false.

2.5.2. Response

How are we to respond to this problem? Biting the epiphenomenalist bullet by denying PP3 is not viable. We saw why epiphenomenalism is unacceptable in Chapter 1 (Section 4.1). Furthermore, accepting epiphenomenalism would put NRIH in a poor dialectical position. NRIH is supposed to offer an alternative to Primitivism. It cannot fend off its Primitivist competitors by citing their unacceptable epiphenomenalist commitments, but then accept similar commitments further down the line. NRIH promised to attribute phenomenal consciousness an appropriate causal status, and should only be supported if it can fulfil that promise.

Denying PP1 has got to be the way forward. NRIH must maintain that our qualitative awareness does make a substantial functional difference, thus allowing it to serve a purpose in our mental economy and to plausibly have an evolutionary origin. As with all of the solutions to the problems discussed, the task is not to present the actual function of consciousness, but rather to show that it is possible for consciousness to have some such function. Consequently, there is no obligation to explain precisely what it is that we can do that a consciousness-free counter-part cannot. Instead, I think we can undermine PP1 by attending to the place of inscrutables in nature.

Fundamental entities have causal powers, and the causal goings-on at higher levels of reality are fully determined by the causal powers instantiated at the fundamental level. What the discussion of our current conception of such entities revealed is that we have no grip on *why* fundamental entities have the dispositions they do. Science identifies entities by their dispositions and maps out the systematic causal relations in which they stand. On the most plausible account of inscrutables, it is the hidden intrinsic nature of these entities that determines their causal powers.

How does this reminder help us to defend NRIH? The notion of inscrutable-inverts only seems plausible if we assume that inscrutables can be re-arranged without this showing up on the level of causal powers. This is simply false. It only *seems* plausible because we have no conception of inscrutables. We know the causal characteristics of things from the outside, and presume that how those entities are on the inside (as it were) can be altered without making any difference to their outward manifestations. We cannot really imagine inscrutables being swapped around because we cannot really imagine inscrutables *at all*. NRIH can claim that if we did have a conception of inscrutables, we would see that they are bound by necessity to the functional economy of the brain. There is not an intrinsic property/causal power property dualism in play. There is thus no justification for the conclusion that one's phenomenal qualities could be inverted or otherwise re-arranged without making a functional difference. We are not in a position to identify what that functional difference would be, but that is beside the point.

Dealing with the pseudo-zombie case is a bit more complicated since it does not involve alterations on the level of inscrutables, but rather alterations on the level of our representation. Nevertheless, our response should be similar. We cannot assume that a being who only represents structural properties (whether of the world or of its own mental states) would have all the same functional capacities as us. A commitment to the existence of inscrutables is a commitment to the real metaphysics of causation being beyond our grip. Assuming, as I think we must, that representation has at least something to do with causal relations, we should also display humility when speculating about possible variations in representational states. We are not in a position to conclude that a counter-part who does not represent their own inscrutables (in the ways necessary for qualitative awareness), could be a (near) functional duplicate of ourselves. Perhaps appropriately representing our inscrutables is integral to the functional economy of our minds. Again, it only *seems* like this is not the case because we are looking at brain processes from the outside. Given our proposed ignorance of the metaphysical foundation of such causal processes, it will inevitably appear that no extra properties are integral to that process occurring, but we have already shown why this appearance cannot be trusted.

Of all the responses to the problems I have put forward, the response to the Purpose Problem makes the boldest appeal to our ignorance. This may be unsatisfying, but such dissatisfaction is something we could have *predicted* as soon as we concluded that dispositions are grounded in inscrutables. So long as our argument for that conclusion is sound, we have all the justification we need. We are not introducing any new kind of ignorance here – rather, we are just working out the full implications of the established ignorance claim. Nevertheless, it will be worthwhile to make some more positive claims about why the proposed representational states occur.

The representation of our own inscrutables may appear purposeless, but I think we can get some grip on its utility. Being responsive to the world is of enormous importance to our mental economy and is obviously the kind of thing that could have an evolutionary origin. We can understand this responsiveness in terms of *information processes*. For instance, we are responsive to certain spectral-reflectance's because the relevant wavelengths affect our sense-organs, which in turn affect certain neurons. This neural state *carries information* about the world: it is a reliable (though fallible) sign that there is an object with the relevant spectral reflectance before us. However, our commitment to inscrutables means that there must be *more* to such information processes than we can currently understand. When we are affected by the world, our causal properties change. Where our causal properties change, there will be a change on the level of intrinsic properties instantiated within us. Consequently, being responsive to properties in the world will involve a correlation between our being presented with that property and certain alterations on the level of our inscrutables. In other words, our inscrutable properties *carry information* about the world. From a third-person perspective, we can only access the outward manifestations of such information-laden correlations, but we can be sure that they are there.

In light of this, we can make some sense of why we would end up representing our own inscrutables. A change among our inscrutables is a mark that the world impresses upon the slate of our mind, and tracking those marks is a way of tracking how things are in the world. The potential utility of such tracking abilities is clear. The next question is why we misrepresent these information-laden inscrutables as externally located rather than as properties of the brain. The answer is that this is

plausibly a *useful* misrepresentation. We need to know how the world is, so representing our inscrutables in a way that *projects* them onto external objects is far more convenient than an inefficient two-tier representation. As Jakab (2003) proposes, such projection is evolution's way round the laborious project of having to assign worldly correlates to each internal signifier. Our projective awareness of our own inscrutable properties is thus a natural consequence of the simple fact that we are responsive to the world in virtue of the differences it makes to us. There is a great deal more that could be said about this, but hopefully we have done enough to undermine the Purpose Problem. Any sense that our representation of inscrutables is without utility is plausibly a reflection of our ignorance. This defuses the worry that NRIH is committed to qualitative awareness being pointless.

SECTION 3

NRIH AND THE PROBLEM OF CONSCIOUSNESS

The Core Thesis of NRIH is that a mental state is a phenomenal state in virtue of suitably representing itself, and is the type of phenomenal state it is in virtue of the unconceived inscrutable properties that implement it. I have now defended this thesis against a number of potential threats. In the process, I have added five further details to NRIH's account of consciousness. The claims are:

> i) That the inscrutable properties doing the explanatory work must be properties of the self-representing state M (thus avoiding the Receptivity Problem).
>
> ii) That M's inscrutables must be represented by M (thus avoiding the Content Problem), though this representation is 'projective'.
>
> iii) That M's representation of its own inscrutables is 'confused' (thus avoiding the Qualitative Character Problem).
>
> iv) That the structural features of a phenomenal state are determined by the content of M, not the actual structure of our inscrutables (thus avoiding the Structural Divergence Problem).

v) That the causal status of M's inscrutables is such that they are integral to M's having the functional profile it has (thus avoiding the Purpose Problem).

We have a proposal about the metaphysical basis of consciousness – about how phenomenal states can arise from physical states. The final task is to explain how this proposal confronts the Problem of Consciousness. In Section 3.1 I offer NRIH's response to the problem. In Section 3.2 I consider the significance of the fact that NRIH's response is bipartite.

3.1. NRIH'S SOLUTION

3.1.1. NRIH and the Criteria of Success

In Chapter 1 (Section 1.3) I introduced the following question: *is the phenomenal onticly dependent on the physical, or onticly independent of the physical?* NRIH answers that the phenomenal is onticly dependent on the physical. Of course, it is this question that leads to the Problem of Consciousness, which I formulated as follows: *there are persuasive reasons to believe that the phenomenal is onticly independent of the physical, and persuasive reasons to believe that the phenomenal is onticly dependent on the physical* (Chapter 1, Section 4.2). NRIH responds to this problem by denying that there are persuasive reasons to adopt Primitivism. The case for Primitivism rests on the apparent epistemic gap between the physical and the phenomenal. NRIH acknowledges this appearance, and explains why it has a grip upon us. Ultimately though, it denies that the epistemic gap is genuine.

In Chapter 2 I identified three criteria that a defensible response to the Problem of Consciousness must satisfy. NRIH holds that phenomenal states are nothing over and above physical states, thus satisfying the Physicalist Criterion. NRIH does not achieve this defence of Physicalism by denying the manifest reality of phenomenal consciousness, so satisfies the Phenomenal Realism Criterion (Chapter 2, Section 1.2). Furthermore, NRIH does not achieve this defence of Physicalism by claiming that the entailment from the physical to phenomenal is a brute a posteriori necessity. Rather, it holds that the psychophysical conditional is knowable a priori to

an appropriately informed subject, thus satisfying the A Priori Entailment Criterion (Chapter 2, Section 2.3). Failure to satisfy these criteria led me to reject Primitivism, Eliminative Type-A Physicalism, and Type-B Physicalism respectively. NRIH is a non-standard form of Type-A Physicalism, and succeeds in satisfying all three of the criteria of success I established. This sets it apart from all familiar attempts to deal with the problem.

There are some further conditions of success that we must also consider. In Chapter 4 I introduced two conditions on EV: the Relevance Condition and the Integration Condition. As a variant on EV, NRIH must be able to satisfy those two conditions. The Relevance Condition required positive reason to believe that unconceived physical properties could evade the a priori obstacles to Physicalism, namely the –tivity and –trinsicality gaps. I will explain how NRIH deals with each of these gaps shortly. The important point here is that NRIH's appeal to inscrutables satisfies this condition. As I argued in Chapter 4, positing inscrutables overcomes the –trinsicality gap. Though positing inscrutables does *not* plausibly avoid the –tivity gap, NRIH does not require them to. The –tivity gap is meant to be dealt with by the Self-Representationalist component of this hybrid proposal. As such, inscrutables are relevant to the specific explanatory task that NRIH attributes them, so NRIH's appeal to inscrutables satisfies the Relevance Condition.

The Integration Condition required positive evidence of a blind-spot in our current conception of the physical that could plausibly be occupied by the proposed unconceived properties. In Chapter 4 (Section 4.3) I showed how the appeal to inscrutables satisfied this condition. NRIH can deploy precisely the same arguments to justify its appeal to inscrutables, and so satisfies the Integration Condition. By combining RIH with Self-Representationalism, we have formed an attenuated version of EV that is capable of meeting both conditions. NRIH is plausibly the best possible version of EV: if you are going to adopt any version of the ignorance hypothesis, you had better adopt the Neo-Russellian Ignorance Hypothesis, or your prospects of satisfying the two conditions on EV are dim.

3.1.2. NRIH and the Epistemic Gap

The case for Primitivism is premised on the apparent epistemic gap between the physical and the phenomenal. NRIH denies that there is such a gap. It holds that for subjects such as ourselves, it is natural to believe that such a gap exists. However, for an ideal subject the psychophysical conditional would be knowable a priori. Of course, we have seen that the epistemic gap is best understood as a composite of the –trinsicality and –tivity gaps. NRIH offers a different response to each of these gaps.

The –trinsicality gap is symptomatic of our limited conception of the physical world. The physical-as-we-know-it is purely structural, yet phenomenal qualities are (absolutely) intrinsic properties. Intrinsic properties cannot be accounted for in structural terms; this is a conceptual principle that NRIH can respect. However, we have compelling reasons to believe that there are *intrinsic* physical properties beyond our current conception. Our ignorance makes it appear that *no* physical property could be suited to the explanation of phenomenal qualities, but having identified our ignorance, we know that this appearance should not be trusted.

The –tivity gap is the apparent conceptual gap between objective physical states and subjective phenomenal states. In contrast to its response to the –trinsicality gap, NRIH does not accept this as a conceptual principle. Once we appreciate the intentional nature of subjectivity, we will see that objective physical states could be the *vehicle* of phenomenal representations. This leaves us free to claim that subjective states are onticly dependent on objective properties. There may remain a deep intuition that subjectivity involves something *more* than the performance of an appropriate vehicular role by objective properties, but this is plausibly symptomatic of a further cognitive error. When objective properties perform the right role, they implement a self-representational state of subjective awareness. But the non-causal nature of our epistemic access to our own subjective states makes it wrongly appear to us that subjectivity has a nature that transcends the performance of any such role. Again, this appearance should not be trusted.

3.1.3. NRIH and the Conceivability Argument

The case for Primitivism is normally presented in terms of the Conceivability Argument (CA) and Knowledge Argument (KA). I have argued that the plausibility of these arguments depends on the plausibility of the –tivity and –trinsicality gaps, each of which NRIH addresses. Nevertheless, it is worth considering exactly how NRIH responds to CA and KA.

CA is based on the claim that we can conceive of replicas of ourselves that are like us in all physical respects, but who differ from us phenomenally. Specifically, we can conceive of *zombie* replicas who have no phenomenal states at all, and *invert* twins whose phenomenal states are qualitatively inverted relative to our own. NRIH's response to this is that we cannot really conceive of such twins. It may *seem* that we conceive of them if we suffer from 'proposition confusion'. Here we do conceive of some genuine possibility, but misidentify what we have imagined. We *can* conceive of a being with the same *structural* profile as ourselves – like us in all functional respects and in all physical respects captured by our current conception of the physical – but who differs from us phenomenally. However, to imagine a structural twin is not to imagine a complete physical twin. There are intrinsic physical properties to be factored in of which we have no conception.

These unknown intrinsic properties are relevant in two ways. First, the fact we have no conception of them means we lack the conceptual resources with which to conceive of a complete physical replica. Consequently, we cannot get as far as conceiving of that replica differing from us phenomenally. Second, NRIH suggests we have reason to believe that *if we did* have a conception of those hidden properties, we would *not* be able to conceive of genuine physical replicas who differ from us phenomenally. For instance, we would be unable to conceive of a being whose phenomenal qualities are inverted relative to our own without imagining appropriate changes on the level of their intrinsic physical properties.

It could be argued that even if we *did* have concepts for inscrutables, zombies would remain conceivable, because it is always conceivable for any collection of objective properties to be unaccompanied by subjective awareness. Self-

Representationalism denies this. With a proper understanding of the physical explanation of intentional properties, and of the intentional explanation of subjectivity, zombie twins will be inconceivable.

3.1.4. NRIH and the Knowledge Argument

KA is based on the intuition that Mary would learn something new on escaping her room. Since she already has complete physical knowledge, what she learns must be a non-physical fact. NRIH responds to this by suggesting that our intuitions about this case are manifestations of our limited conception of the physical, and so of our skewed understanding of what Mary would learn within her room. With full knowledge of the inscrutable properties of the brain, and their place in our representational-functional economy, Mary could deduce that seeing a tomato would involve *that* phenomenal quality. This deduction is only possible if Mary already has a *concept* of phenomenal redness before leaving her room. She would not be able to infer what redness is like from her physical knowledge unless she had previously acquired a concept of phenomenal redness by experiencing if for herself. However, as I argued in Chapter 1 (Section 2.3.1), the only plausible version of KA grants that Mary has this concept *before* leaving her room. The Primitivist claim is that she learns, on leaving her room, that this phenomenal quality is associated with that physical brain state rather than some other. NRIH denies this.

A possibility explored in Section 3 of this chapter also has ramifications for KA. The entailment from inscrutables to phenomenal qualities might involve a type of explanation beyond our current grasp. If this were so, we would have the intuition that Mary could never deduce qualitative facts from non-qualitative facts, but this could be a reflection of our ignorance of an *explanation type* available to Mary but not to us. One speculation along these lines is that inscrutables have propensities to ‘fuse’ such that knowledge of inscrutables and their propensities would allow one to deduce which inscrutables are responsible for which phenomenal qualities. In this scenario, our ignorance of that variety of ontic dependence sends our intuitions awry, but for Mary the path from inscrutables to phenomenal qualities is perfectly clear.

3.2. A CONFLUENCE OF ILLUSIONS?

According to NRIH, two factors make consciousness appear physically inexplicable. If one embraced a Self-Representationalist theory of subjectivity in isolation, the qualitative character of consciousness would remain mysterious and draw one towards Primitivism. Similarly, if one embraced the Russellian approach to qualitative character in isolation, the subjectivity of consciousness would remain mysterious and again draw one towards Primitivism. The subjective character and qualitative character of phenomenal states appear physically inexplicable for *two distinct reasons*, each of which would be sufficient to generate the impression of an epistemic gap on their own.

Kriegel (2009) discusses the possibility that '…several distinct properties suffice individually to generate the mystery in an overdetermining fashion' (pp.6-7). He goes on to judge that though this is not impossible, it is quite implausible (2009, p.7). This captures a worry that many might have when presented with NRIH. Is it really plausible that nature has conspired to make one phenomenon appear inexplicable to us twice over? This does not constitute a knock-down objection to NRIH, but it is a worry that is worth considering. There are two ways of assuaging this worry. First, to suggest that the overdetermination of the mystery is not so implausible. Second, to show that NRIH does not posit two distinct sources of illusion after all, but rather two aspects of a single piece of cognitive trickery.

3.2.1. An Overdetermined Illusion

Two considerations indicate that the overdetermination of our sense of mystery is not so implausible. First, I have consistently put significant weight on the distinction between the –tivity and –trinsicality gaps. If we maintain, as I think we should, that any serious formulation of the epistemic gap consists in a combination of these two gaps, it will be no surprise that there are two distinct illusions in play. One illusion would generate the first apparent gap, and the other would generate the second.

Second, the limitations of certain existing responses to the epistemic gap can be diagnosed in terms of their failure to recognise that the illusion of consciousness being physically inexplicable is overdetermined. Those responses correctly identify *one* of the sources of illusion, but then *exaggerate* its significance, driven by the assumption that our sense of mystery has a single origin.

Strawson and others identify that we have a limited conception of matter, and that the conviction that our conception of matter is relevantly complete makes consciousness appear physically inexplicable. According to NRIH, this much is true. However, if it is assumed that our sense of mystery has just one source, then the unknown aspects of matter had better be *entirely* responsible for consciousness. This leads to the panphenomenalist claim that the intrinsic nature of matter is inherently experiential. Of course, it is this panphenomenalist commitment that makes Strawson's position implausible.

Contrast this with the Representationalist approach to our sense of mystery. Self-Representationalism, and its HOR-theory cousins, claim that consciousness only seems physically inexplicable because we have failed to appreciate that consciousness is an intentional matter. Again NRIH agrees, at least where the –tivity gap is concerned. However, if it is assumed that our sense of mystery has a single origin, then Representationalism should be able to deal with the whole epistemic gap single-handedly. This encourages misguided Representationalist accounts of the qualitative character of phenomenal states. Indeed, Kriegel offers just such an account (2009, Chapter 3).[20]

The insights at the heart of NRIH have not sprung from nowhere. They are connected intimately to a great deal of work already done on the Problem of Consciousness. We can explain the strengths of those existing positions in terms of their correct identification of one of the two sources of mystery, and the failings of those positions in terms of the assumption that *their* source is the *only* source. The

[20] Interestingly, Kriegel concedes that the greatest objection to his account of qualitative character is that it makes phenomenal qualities dispositional properties, though they appear to us to be occurrent monadic properties (2009, p.95). In other words, he wrongly construes phenomenal qualities as *structural* properties rather than *intrinsic* properties, which I have argued is the fatal flaw of all Representationalist accounts of qualitative character (Chapter 5, Section 2).

partial success but ultimate failure of existing positions is an interesting datum, and the claim that the appearance of inexplicability is overdetermined is a plausible explanation of that datum. In this sense, the unexpected overdetermination of our sense of mystery is partly responsible for the Problem of Consciousness being so difficult to solve. A key virtue of NRIH is that it extracts what is best about existing positions whilst avoiding their mistakes.

3.2.2. One Illusion, Two Manifestations

Another reason for not being worried by NRIH's commitment to the overdetermination of the appearance of an epistemic gap is that the two proposed sources plausibly have a deeper *common* source. The source of the apparent –trinsicality gap is that science only reveals the causal profile of physical entities, never their intrinsic nature. The fact that consciousness is the only context in which inscrutables are manifest to us leads us towards a false picture of the physical world as devoid of intrinsic properties: a picture that cannot accommodate phenomenal qualities. The source of the apparent –tivity gap is that consciousness is the only context in which we know a state directly through *being in it* rather than knowing it via its causal profile. The fact that consciousness is the only context in which we have non-causal epistemic access to something leads us towards a false picture of subjective awareness as something over and above the performance of an appropriate role.

What this means is that *both* apparent gaps are manifestations of the within/without epistemic duality. If only we had access to the *inside* of external objects – their intrinsic nature – the qualitative character of conscious states would not seem inexplicable. If only we accessed our subjective awareness from the *outside* – through its causal profile – its instantiation would not seem to require anything more than the performance of an appropriate causal role by objective properties. Put another way, the receptive nature of our knowledge of the physical world makes qualitative character appear physically inexplicable, and the distinctively *non*-receptive nature of phenomenal self-knowledge makes subjectivity appear physically inexplicable.

On this view, the assumption that our epistemic access to things discloses their full nature is responsible for *both* apparent gaps. The assumption that the causal powers of external objects exhausts their nature, when in fact they have an absolutely intrinsic aspect, is what generates the apparent –trinsicality gap. The assumption that our direct access to our subjective awareness reveals its full nature, when in fact it is constituted by the performance of an appropriate vehicular role, is what generates the –tivity gap. Our failure to recognise that both modes of epistemic access are limited, each revealing just one aspect of reality, is the single source of the appearance of an epistemic gap between the physical and the phenomenal.

This suggests that NRIH is not committed to an implausible confluence of illusions. It does not claim that our sense that consciousness is physically inexplicable is overdetermined. Rather, the illusion of inexplicability has one route with two very different manifestations. This conclusion is quite compatible with the points made in Section 3.2.1 about the failure of existing positions to acknowledge the two-part structure of the epistemic gap. That still stands, but must be understood in terms of two *aspects* of the apparent gap rather than two genuinely distinct sources of illusion. This discussion should also assuage worries, raised at the beginning of the chapter, about RIH and Self-Representationalism being an awkward pairing. We can now see that these two attempts to solve the problem each concern different sides of the same coin. It is clear that they are natural partners, and that together they form an integrated hybrid response to the Problem of Consciousness.

CONCLUSION

NRIH undermines the epistemic gap at the heart of the Problem of Consciousness whilst offering a powerful account of why that apparent gap seems so compelling. Central to its success is the proposal that the epistemic gap is really a composite of the –tivity and –trinsicality gaps. This division has proved itself to be invaluable in the evaluation of attempted responses to the Problem of Consciousness. NRIH offers a non-standard account of the metaphysical status of consciousness that avoids the key failings of standard accounts. It also offers the best possible way of deploying the key insights of EV and of Representationalism. Further work in a number of areas would

serve to reinforce the case for NRIH. For instance, I have taken it that representation can be accounted for in physical terms, though the formulation of such an account is a challenging on-going project. Overall, NRIH is a plausible metaphysical model of the phenomenal that provides a powerful and distinctive response to the Problem of Consciousness, and which offers a serious alternative to the entrenched approaches to that problem.

BIBLIOGRAPHY

Allais, L. (2006) 'Intrinsic Natures: A Critique of Langton on Kant', *Philosophy and Phenomenological Research*, 23(1), 143-169

Alter, T. (2007) 'Does Representationalism Undermine the Knowledge Argument?' in Alter, T. & Walter, S. (eds.) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* Oxford: OUP, pp.65-76

Alter, T. (2009) 'Does the Ignorance Hypothesis Undermine the Conceivability and Knowledge Arguments?', *Philosophy and Phenomenological Research*, 79(3), 756-765

Banks, E.C. (2010) 'Neutral monism reconsidered', *Philosophical Psychology*, 23(2), 173-187

Bayne, T. (2001) 'Chalmers on the Justification of Phenomenal Judgements', *Philosophy and Phenomenological Research*, 62(2), 407-419

Bennett, K. (2009) 'What you don't know can hurt you', *Philosophy and Phenomenological Research*, 79(3), 766-774

Bigelow, J. Ellis, B. & Lierse, C. (1992), 'The World as One of a Kind', *The British Journal for the Philosophy of Science*, 43(3), 371- 388

Bird, A. (2007) 'The Regress of Pure Powers?', *Philosophical Quarterly*, 57(229), 513-534

Blackburn, S. (1990) 'Filling in Space', *Analysis,* 50, 62-65

Block, N. & Stalnaker, R. (2002) 'Conceptual Analysis, Dualism, and the Explanatory Gap' in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.371-393

Block, N. (1980) 'Troubles with Functionalism', in Block (ed.) *Readings in the Philosophy of Psychology Vol.1*. MA: Harvard University Press. pp.263-303

Block, N. (1995) 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences,* 18, 227–287

Block, N. (2002) 'Concepts of Consciousness', in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.206-218

Boghossian, P. & Velleman, D. (1991) 'Physicalist Theories of Color' in Byrne & Hilbert (eds.) *Readings on Color Vol.1:The Philosophy of Color.* Cambridge, MA: MIT Press, pp.105-136

Bolender, J. (2001) 'An Argument for Idealism', *Journal of Consciousness Studies*, 8(4), 37-61

Boutel, A. (forthcoming) 'How to be a Type-C Physicalist', *Philosophical Studies*

Braddon-Mitchell, D. & Jackson, F. (1996) *Philosophy of Mind and Cognition*. Oxford: Blackwell

Broad, C.D. (1925) *The Mind and Its Place in Nature.* London: Routledge & Kegan Paul

Brueckner, A.L. & Beroukhim, E. (2003) 'McGinn on Consciousness and the Mind-Body Problem', in Smith & Jokic (eds.), *Consciousness: New Philosophical Perspectives*. Oxford: OUP, pp.396-408

Burge, T. (1988) 'Individualism and Self-Knowledge', *Journal of Philosophy*, 85: 649-663

Byrne, A. (1999) 'Cosmic Hermeneutics', *Philosophical Perspectives*, 13, 347-383

Byrne, A. (2001) 'Intentionalism Defended', *Philosophical Review,* 110, 199-239

Cao, T.Y. (2003) 'Can we dissolve physical entities into mathematical structure?', *Synthese*, 136, 51–71

Chakravartty, A. (2003) 'The Structuralist Conception of Objects', *Philosophy of Science* 70, 867–878

Chakravartty, A. (2004) 'Structuralism as a Form of Scientific Realism', *International Studies in Philosophy of Science*, 18, 151-171

Chalmers, D. & Jackson, F. (2001) 'Conceptual Analysis and Reductive Explanation', *Philosophical Review*, 110, 315-61

Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: OUP

Chalmers, D. (2002) 'Consciousness and its place in nature' in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.247-272

Chalmers, D. (2003) 'The Content and Epistemology of Phenomenal Belief' in Smith & Jokic (eds.) *Consciousness: New Philosophical Perspectives*. Oxford: OUP, pp.220-272

Chalmers, D. (2004) 'The Representational Character of Experience' in Leiter (ed.) *The Future of Philosophy*. Oxford: OUP, pp.153-181

Chalmers, D. (2007) 'Phenomenal Concepts and the Explanatory Gap' in Alter & Walter (eds.) *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: OUP, pp.167-194

Chomsky, N. (2009) 'The Mysteries of Nature: How Deeply Hidden?', *The Journal of Philosophy*, 106(4), 167-200

Churchland, P. S. (1996) 'The Hornswoggle Problem', *Journal of Consciousness Studies*, 3(5-6), 402-408

Clifford, W.K. (1878) 'On the Nature of Things-in-Themselves', *Mind*, 3(9), 57-67

Coleman, S. (2006) *'*Being Realistic: Why Physicalism May Entail Panexperientialism', *Journal of Consciousness Studies, 13(10–11), 40-52*

Coleman, S. (2007) 'Review of Rosenberg *A Place for Consciousness* and Stoljar *Ignorance and Imagination*', *Philosophical Psychology*, 20(6), 826-833

Coleman, S. (2008), 'Mind Under Matter' in Skrbina (ed.) *Mind that Abides*. Amsterdam: Benjamins, pp.83-108

Coleman, S. (2012) 'Mental Chemistry: Combination for Panpsychists', *Dialectica*, 66(1), 137-166

Crane, T. & Mellor, D.H. (1990) 'There is no Question of Physicalism', *Mind*, 99(394), 185-206

Crane, T. (2005) 'The Problem of Perception', *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/perception-problem/> (Last accessed 20/05/2012)

Crane, T. (2007) 'Intentionalism', Available from: <http://web.mac.com/cranetim/Tims_website/Online_papers_files/Intentionalism.pdf> (Last accessed 13/06/2012)

Cruse, P. (2004) 'Scientific realism, Ramsey sentences and the reference of theoretical terms', *International Studies in the Philosophy of Science,* 18(1-2), 133-149

Dauer, F.W. (2001), 'McGinn's Materialism and epiphenomenalism', *Analysis*, 61(2), 136-39

Davies, M. (2008) 'Consciousness and Explanation' Available from: <http://www.philosophy.ox.ac.uk/__data/assets/pdf_file/0006/2112/Consciousness.pdf> (Last accessed 13/06/2012)

Dennett, D.C. (1991a) *Consciousness Explained*. Boston MA: Little-Brown

Dennett, D.C. (1991b) 'Review of C. McGinn, The Problem of Consciousness', *The Times Literary Supplement*, 10 May

Dennett, D.C. (1996) 'Facing Backwards on the Problem of Consciousness, *Journal of Consciousness Studies*, 3(1), 4-6

Dretske, F. (1986) 'Misrepresentation' in Bogdan (ed.) *Belief: Form, Content and Function*. Oxford: OUP, pp.17-36

Dretske, F. (2003) 'Experience as Representation', *Philosophical Issues*, 13, 67-82

Egan, A. (2010) 'Projectivism without error' in Nanay (ed.) *Perceiving the World*. Oxford: OUP, pp.68-96

Ellis, B. & Lierse, C. (1994) 'Dispositional Essentialism', *Australasian Journal of Philosophy,* 72(1), 27-45

Esfeld, M. & Lam, V. (2008) 'Moderate Structural Realism About Space Time', *Synthese,* 160, 27-46

Farrell, B.A. (1950) 'Experience', *Mind*, 59, 170-198

Feigl, H. (2002) 'The "Mental" and the "Physical"', reprinted in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.247-272

Feser, E. (2001) 'Qualia: Irreducibly Subjective but not Intrinsic', *Journal of Consciousness Studies*, 8(8), 68-72

Flanagan, O. (1992) *Consciousness Reconsidered*. Cambridge MA: MIT Press

Foster, J. (2008) *A World for Us: The Case for Phenomenalistic Idealism.* Oxford: OUP

Francescotti, R.M. (1999) 'How to Define Intrinsic Properties', *Noûs*, 33, 590-609

French, S. & Ladyman, J. (2003), 'Remodelling Structural Realism: Quantum Physics and the Metaphysics of Structure', *Synthese*, 136, 31-56

Gennaro, R.J. (1996), *Consciousness and Self-Consciousness*. Amsterdam: John Benjamins

Gennaro, R.J. (2004) 'Higher-Order Thoughts, Animal Consciousness, and Misrepresentation' in Gennaro (ed.) *Higher-Order Theories of Consciousness.* Amsterdam: John Benjamins, pp.1-16

Gertler, B. (2009) 'The Role of Ignorance in the Problem of Consciousness', *Noûs*, 43(2), 378-393

Goff, P. (2006) 'Experiences Don't Sum', Journal of Consciousness Studies, 13(10–11), 53-61

Goldman, A. (1993) 'Consciousness, Folk Psychology and Cognitive Science', *Consciousness and Cognition*, 2, 364-383

Güzeldere, G. (1997) 'The Many Faces of Consciousness: A Field Guide', in Block, Flanagan & Güzeldere (eds.) *The Nature of Consciousness: Philosophical Debates*. Cambridge MA: MIT Press, pp. 1-68

Hardcastle, V.G. (2008) 'Review of Stoljar, Ignorance and Imagination'*, Philosophical Books*, 49(3), 274-275

Harris, R. (2010) 'How to define extrinsic properties', *Axiomathes,* 20, 461-478

Heil, J. (2005) 'Dispositions', *Synthese*, 144, 343-56

Hellie, B. (2007) 'Higher-order intentionality and higher-order acquaintance', *Philosophical Studies*, 134, 289-324

Hill, C. (1997) 'Imaginability, Conceivability, Possibility, and the Mind-Body Problem', *Philosophical Studies*, 87, 61-85

Hill, C. (2004) 'Ouch! An Essay on Pain' in Gennaro (ed.) *Higher-Order Theories of Consciousness.* Amsterdam: John Benjamins, pp.339-362

Hohwy, J. (2005) 'Explanation and Two Conceptions of the Physical', *Erkenntnis,* 62, 71–89

Holman, E.L. (2008) 'Panpsychism, Physicalism, Neutral Monism and the Russellian Theory of Mind', *Journal of Consciousness Studies*, 15(5), 48-67

Holton, R. (1999) 'Dispositions All the Way Round', *Analysis*, 59, 9-14

Horgan, T. & Tienson, J. (2002) 'The intentionality of phenomenology and the phenomenology of intentionality', in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.520-533

Horgan, T. (1993) 'From Supervenience to Superdupervenience: Meeting the Demands of the Natural World', *Mind*, 102, 555-586

Humberstone, I.L. (1996) 'Intrinsic/Extrinsic', *Synthese,* 108, 205-267

Hume, D. (1739) *A Treatise of Human Nature*. Norton ed. (2006) Oxford: Clarendon

Hutton, D.D. (2009) 'Mental Representation and Consciousness' in Banks (ed.) *The Encyclopedia of Consciousness*. Oxford: Academic Press, Vol. 2, pp.19-33

Jackson, F. (1982) 'Epiphenomenal Qualia', *Philosophical Quarterly,* 32, 127-36

Jackson, F. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Clarendon Press: Oxford

Jackson, F. (2007) 'The Knowledge Argument, Diaphanousness, Representationalism' in Alter & Walter eds. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism.* Oxford: OUP, pp.52-64

Jakab, Z. (2003) 'Phenomenal Projection', *Psyche*, Vol.9

James, W. (1890), *The Principles of Psychology*. New York: Henry Holt

Judisch, N. (2008) 'Why "Non-mental" Won't Work: On Hempel's Dilemma and the Characterization of the "Physical"', *Philosophical Studies*, 104, 299-318

Kant, I. (1781/1787) *The Critique of Pure Reason*. Pluhar trans. (1996), Indianapolis: Hackett

Kim, J. (1982) 'Psychophysical supervenience', *Philosophical Studies*, 41(1), 51-70

Kim, J. (1989) 'The Myth of Nonreductive Materialism', *Proceedings and Addresses of the American Philosophical Association*, 63(3), 31-47

Kim, J. (2002) 'The Many Problems of Mental Causation', in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.170-178

Kind, A. (2007) 'Restrictions on Representationalism', *Philosophical Studies*, 134, 405-427

Kirk, R. (1991) 'Why shouldn't we be able to solve the mind-body problem?', *Analysis,* 15(1), 17-23

Kirk, R. (2006) 'Physicalism and Strict Implication', *Synthese*, 151, 523-536

Kriegel, U. (2002) 'PANIC theory and the prospects for a representational theory of phenomenal consciousness', *Philosophical Psychology*, 15(1), 55-64

Kriegel, U. (2003) 'The New Mysterianism and the Thesis of Cognitive Closure', *Acta Analytica,* 18(30/31), 177-191

Kriegel, U. (2005) 'Naturalizing Subjective Character', *Philosophy and Phenomenological Research*, 71, 23-57

Kriegel, U. (2009) *Subjective Consciousness: A Self-Representational Theory*. Oxford: OUP

Kripke, S. (1980) *Naming and Necessity.* Cambridge, MA: Harvard University Press

Kukla, A. (1995) 'Mystery, mind and materialism', *Philosophical Psychology*, 8(3), 255-264

Ladyman, J. & Ross, D. (with Spurrett, D. and Collier, J.) (2007) *Every Thing Must Go: Metaphysics Naturalised*. Oxford: OUP

Ladyman, J. (2007) 'On the Identity and Diversity of Objects in a Structure', *Proceedings of the Aristotelian Society Supplementary Volume*, 81, 23-43

Langton, R. (1998) *Kantian Humility: Our Ignorance of Things in Themselves*. Oxford: Clarendon Press

Langton, R. (2004) 'Elusive Knowledge of Things in Themselves', *Australasian Journal of Philosophy*, 82(1), 129-136

Levine, J. (2001) *Purple Haze: The Puzzle of Consciousness*. Oxford: OUP

Levine, J. (2002) 'Materialism and Qualia: The Explanatory Gap' in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.354-361

Levine, J. (2006) 'Conscious Awareness and (Self-) Representation', in Kriegel & Williford (eds.) *Self-Representational Approaches to Consciousness*. Cambridge MA: MIT Press, pp.173-198

Lewis, D. (1970) 'How to define theoretical terms', *The Journal of Philosophy*, 67(13), 427-446

Lewis, D. (1983) 'Extrinsic Properties', *Philosophical Studies*, 44, 197-200

Lewis, D. (2009) 'Ramseyan Humility', in Nola & Braddon-Mitchell (eds.) *Conceptual Analysis and Philosophical Naturalism*. Cambridge MA: MIT Press, pp.203-222

Lipton, P. & Worrall, J. (2000) 'Tracking Track Records', *Proceedings of the Aristotelian Society Supplementary Volumes*, (74), 179-205 + 207-223

Loar, B. (1990) 'Phenomenal States', *Philosophical Perspectives,* 4, 81-108

Locke, J. (1690) *An Essay Concerning Human Understanding,* Nidditch ed. (1975), Oxford: OUP

Lockwood, M. (1989) *Mind, Brain and the Quantum: The Compound 'I'*. Oxford: Basil Blackwell

Lockwood, M. (1993) 'The Grain Problem' in Robinson (ed.), *Objections to Physicalism*. Oxford: Clarendon Press, pp.271-291

Look, B.C. (2008) 'Gottfried Wilhelm Leibniz', in Zalta (ed.) *The Stanford Encyclopedia of Philosophy* URL = <http://plato.stanford.edu/archives/fall2008/entries/leibniz/> (Last Accessed 13/06/2012)

Lycan, W.G. (1996) *Consciousness and Experience*. Cambridge, MA: MIT Press

Lycan, W.G. (2001) 'A Simple Argument for a Higher-Order Representation Theory of Consciousness' *Analysis*, 61(1), 3-4

Lycan, W.G. (2004) 'The Superiority of HOP to HOT', in Gennaro (ed.) *Higher-Order Theories of Consciousness.* Amsterdam: John Benjamins, pp.93-114

Lycan, *W.G.* (2006) 'Resisting ?-ism'*, Journal of Consciousness Studies,* 13(10-11), 65-71

Mach, E. (1886), *The Analysis of Sensations and the Relation of Physical to the Psychical.* Williams & Waterlow trans. (1959), New York: Dover

Martin, C.B. & Heil, J. (1999), 'The Ontological Turn', *Midwest Studies in Philosophy*, 23, 34-60

Maxwell, G. (2002) 'Rigid Designators and Mind-Brain Identity', in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.341-353

McClelland, T. (2012) 'In Defence of Kantian Humility', *Thought*, 1, 62-70

McClelland, T. (forthcoming) 'Review of Pereboom *Consciousness and the Prospects of Physicalism*', *Journal of Consciousness Studies*

McGeer, V. (2003) 'The Trouble with Mary', *Pacific Philosophical Quarterly*, 84, 384-393

McGinn, C. (1989) 'Can we solve the mind-body problem?', *Mind*, 98, 349-66

McGinn, C. (1994) 'Reply to Carol Rovane', *Philosophical Studies,* 76, 169-174

McGinn, C. (2004) *Consciousness and Its Objects.* Oxford: Clarendon Press

McKitrick, J. (2003) 'The Bare Metaphysical Possibility of Bare Dispositions', *Philosophy and Phenomenological Research*, 66(2), 349-369

McKitrick, J. (2006) 'Rosenberg on Causation', *Psyche*, 12(5)

McLaughlin, B.P. (2007) 'On the Limits of A Priori Physicalism' in McLaughlin and Cohen (eds.) *Contemporary Debates in Philosophy of Mind.* Maldon, MA: Blackwell, pp.200-223

Megill, J.L. (2005) 'Locke's mysterianism: On the unsolvability of the mind-body problem', *Locke Studies*, 5, 119-147

Meixner, U. (2004) *The Two Sides of Being: A Reassessment of Psycho-Physical Dualism*. Paderborn: Mentis

Mellor, D.H. (1995) *The Facts of Causation*. London: Routledge

Metzinger, T. (1995) 'Introduction' to Metzinger (ed.) *Conscious Experience*. Schoningh: Imprint Academic

Metzinger, T. (2003) *Being No One.* Cambridge, MA: MIT Press

Montero, B. (1999) 'The Body Problem', *Noûs*, 33(2), 183-200

Montero, B. (2010) 'A Russellian Response to the Structural Argument Against Physicalism', *Journal of Consciousness Studies*, 17(3-4), 70–83

Nagasawa, Y. (2010) 'The Knowledge Argument and Epiphenomenalism', *Erkenntis,* 72, 37-56

Nagel, T. (1974) 'What is it like to be a bat?', *Philosophical Review*, 83, 435-50

Nagel, T. (1986) *The View From Nowhere*. Oxford: OUP

Neander, K. (1998) 'The Division of Phenomenal Labour: A Problem for Representational Theories of Consciousness', *Noûs*, 32(12), 411-434

Papineau, D. (1996) 'Theory-dependent terms', *Philosophy of Science*, 63, 1-21

Papineau, D. (2002) *Thinking About Consciousness*. Oxford: Clarendon

Papineau, D. (2007) 'Review of Daniel Stoljar, Ignorance and Imagination', *Notre Dame Philosophical Reviews,* 4

Papineau, D. (2011) 'What Exactly Is the Explanatory Gap?', *Philosophia*, 39, 5-19

Penrose, R. (1989) *The Emperor's New Mind*. Oxford: OUP

Pereboom, D. (2011) *Consciousness and the Prospects of Physicalism*. New York: OUP

Perry, J. (1979) 'The Problem of the Essential Indexical', *Noûs,* 13(1), 3-21

Polger, T.W. (2008) 'H_2O, "Water", and Transparent Reduction', *Erkenntnis*, 69, 109-30

Prinz, J. (2003) 'Level-headed mysterianism and artificial experience', *Journal of Consciousness Studies*, 10(4-5), 111–32

Psillos, S. (2006) 'The Structure, the Whole Structure and Nothing But the Structure?', *Philosophy of Science*, 73, 560-570

Puryear, S. (2005) 'Was Leibniz Confused About Confusion?' *Leibniz Review*, 15, 95-124

Rey, G. (1997) *Contemporary Philosophy of Mind,* Oxford: Blackwell

Robinson, H. (unpublished) 'Qualia, Qualities and our Conception of the Physical World'

Rosenberg, G. (2004) *A Place for Consciousness: Probing the Deep Structure of the Natural World*. Oxford: OUP

Rosenthal, D.M. (1986) 'Two concepts of consciousness', *Philosophical Studies*, 49, 329-359

Rosenthal, D.M. (2004) 'Varieties of Higher Order View' in Gennaro (ed.) *Higher-Order Theories of Consciousness.* Amsterdam: John Benjamins, pp.17-44

Rovane, C. (1994) 'A comment on McGinn's *The problem of philosophy*', *Philosophical Studies*, 76, 157-68

Russell, B. (1910) 'Knowledge by acquaintance and knowledge by description', *Proceedings of the Aristotelian Society*, 11, 108-128.

Russell, B. (1921) *The Analysis of Mind*. London: George Allen & Unwin Ltd.

Russell, B. (1927), *The Analysis of Matter*. London: George Allen & Unwin Ltd.

Sacks, M. (1994) 'Cognitive Closure and the Limits of Understanding', *Ratio*, 7, 26-42

Saunders, S. (2003) 'Structural realism again', *Synthese*, 136, 127-133

Schopenhauer, A. (1819/1844) *The World as Will and Representation*. Payne trans. (1969), New York: Dover

Seager, W. (1995) 'Consciousness, Information, and Panpsychism', *Journal of Consciousness Studies*, 2, 272-88

Seager, W. (2006) 'The "Intrinsic Nature" Argument for Panpsychism', *Journal of Consciousness Studies*, 13(10-11), 129-145

Seager, W. (2010) 'Panpsychism, Aggregation and Combinatorial Infusion', *Mind and Matter,* 8(2)*,* 167-184

Searle, J.R. (1990) 'Consciousness, explanatory inversion and cognitive science', *Behavioral and Brain Sciences,* 13, 585-642

Shoemaker, S. (1982) 'The Inverted Spectrum', *Journal of Philosophy*, 79(7), 357-381

Shoemaker, S. (1997) 'Causality and Properties' in Mellor & Oliver (eds.), *Properties.* Oxford: OUP, pp.228-255

Smart, J.J.C. (1959) 'Sensations and Brain Processes', *The Philosophical Review*, 68(2), 141-156

Sprigge, T.L.S. (1971) 'Final Causes', *Supplementary Proceedings of the Aristotelian Society*

Spurrett, D. & Papineau, D. (1999) 'A note on the completeness of "physics"', *Analysis*, 59, 25–29

Stoljar, D. & Nagasawa, Y. (2003) 'Introduction' to Ludlow, Stoljar & Nagasawa (eds.) *There's Something About Mary*. Cambridge, MA: MIT Press, pp.1-36

Stoljar, D. (2001) 'Two Conceptions of the Physical', *Philosophy and Phenomenological Research,* 62(2), 253-281

Stoljar, D. (2005) 'Physicalism and Phenomenal Concepts', *Mind and Language,* 20, 469-94

Stoljar, D. (2006) *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness.* Oxford: OUP

Stoljar, D. (2009), 'Response to Alter and Bennett', *Philosophy and Phenomenological Research*, 79(3), 775-784

Stoljar, D. (2010) *Physicalism*. London: Routledge

Strawson, G. (1994) *Mental Reality*. Cambridge, MA: MIT Press

Strawson, G. (2006) 'Realistic Monism: Why Physicalism Entails Panpsychism', Journal of Consciousness Studies, 13(*10-11), pp.3-31*

Strawson, G. (2008) *Real Materialism and Other Essays*. Oxford: OUP

Strawson, G. (2011) 'Radical Self-Awareness', in Siderits, Thompson, & Zahavi (eds.) *Self, No Self?: Perspectives from Analytical, Phenomenological, and Indian Traditions*. Oxford: OUP, pp. 274–307

Trogden, K. (2009) 'Review of Stoljar *Ignorance and Imagination*', *Philosophical Review*, 118(2), 269-273

Tye, M. (1995) *Ten Problems of Consciousness: A Representational Theory of the Phenomenal Mind*. Cambridge, MA: MIT Press

Tye, M. (2000) *Consciousness Color and Content*. Cambridge, MA: MIT Press

Unger, P. (1998) 'The Mystery of the Physical and the Matter of Qualities', *Midwest Studies in Philosophy*, 23, 75-99

Van Cleve, J. (1988), 'Inner States and Outer Relations: Kant and the Case for Monadism', in Hare (ed.) *Doing Philosophy Historically*. Buffalo, NY: Prometheus, pp. 231–47

Van Cleve, J. (2002), 'Receptivity and Our Knowledge of Intrinsic Properties', *Philosophy and Phenomenological Research*, 65(1), 218-237

Van Fraassen, B.C. (2007) 'Structuralism(s) about science: some common problems', *Proceedings of the Aristotelian Society Supplementary Volume,* 81, 45-61

Van Gulick, R. (2004) 'Higher-order global states (HOGS): An alternative higher-order model of consciousness' in Gennaro (ed.) *Higher-Order Theories of Consciousness.* Amsterdam: John Benjamins, pp.67-92

Van Gulick, R. (2006) 'Mirror mirror - is that all?' in Kriegel & Williford (eds.) *Self-Representational Approaches to Consciousness*. Cambridge, MA: MIT Press, pp.11-40

Van Gulick, R. (2011) 'Subjective Consciousness and Self-Representation', *Philosophical Studies*, Available from: <http://www.springerlink.com.ezproxy.sussex.ac.uk/content/4570474672727206/fulltext.html> (Last accessed: 13/06/2012)

Whiteley, C.H. (1990) 'McGinn on the Mind-Body Problem', *Mind*, 99, 394

Witmer, D.G. (2006) 'How to be a (Sort of) A Priori Physicalist', *Philosophical Studies*, 131, 185-225

Worrall, J. (1989) 'Structural realism: The best of both worlds?', *Dialectica,* 43, 99-124

Wright, W. (2003) 'Projectivist Representationalism and Color', *Philosophical Psychology*, 16(4), 515-533

Yablo, S. (2002), 'Mental Causation', in Chalmers (ed.) *Philosophy of Mind: Classical and Contemporary Readings.* Oxford: OUP, pp.179-196